Essays on Causal Inference for Public Policy

Essays on Causal Inference for Public Policy

Abstract

Effective policymaking requires understanding the causal effects of competing proposals. Relevant causal quantities include proposals' expected effect on different groups of recipients, the impact of policies over time, the potential trade-offs between competing objectives, and, ultimately, the optimal policy. This dissertation studies causal inference for public policy, with an emphasis on applications in economic development and education.

The first chapter introduces Bayesian methods for time-varying treatments that commonly arise in economics, health, and education. I present methods that account for dynamic selection on intermediate outcomes and can estimate the causal effect of arbitrary dynamic treatment regimes, recover the optimal regime, and characterize the set of feasible outcomes under different regimes. I demonstrate these methods through an application to optimal student tracking in ninth and tenth grade mathematics. The proposed estimands characterize outcomes, mobility, equity, and efficiency under different tracking regimes.

The second chapter studies regression discontinuity designs with multiple forcing variables. Leading examples include education policies where treatment depends on multiple test scores and spatial treatment discontinuities arising from geographic borders. I give local linear estimators for both the conditional effect along the boundary and the average effect over the boundary. For two-dimensional RD designs, I derive an optimal, data-dependent, bandwidth selection rule for the conditional effect. I demonstrate these methods using a summer school and grade retention example.

The third chapters illustrate the central role of persistence in estimating and interpreting value-added models of learning. Using data from Pakistani public and private schools, I apply dynamic panel methods that address three key empirical challenges: imperfect persistence, unobserved student heterogeneity, and measurement error. After correcting for these difficulties, the estimates suggest that only a fifth to a half of learning persists between grades and that private schools increase average achievement by 0.25 standard deviations each year. In contrast, value-added models that assume perfect persistence yield severely downwardly biased and occasionally wrong-signed estimates of the private school effect.

Acknowledgements

This dissertation was nurtured to fruition by an amazing group of mentors. I am deeply grateful to my committee – Alberto Abadie, Guido Imbens, Dale Jorgenson, and Asim Khwaja – for their constant support and insightful feedback. I am especially indebted to my primary advisor, Guido Imbens, for providing encouragement as I spent an increasing amount of time thinking about empirical methods for policy evaluation. The clarity of Guido's own work remains a model towards which I strive. Asim Khwaja has been a long-time collaborator and friend whose quick mind and love of ideas has made the doctoral journey far richer and more rewarding. Alberto Abadie always provided careful and thoughtful feedback, and I could ask for no better advisor to study policy evaluation. I am thankful to have been taught by and taught under Dale Jorgenson. Always the perfect gentleman, I will remember his sincere interest in all of his students and their work.

My dissertation journey would never have begun without the early encouragement by my undergraduate professors, Tahir Andrabi, Frank Wykoff, and Jay Atlas. Tahir Andrabi introduced me to development economics with ever-present humor and good sense. Frank Wykoff taught me econometrics through the ample use of a red pen. I still recall him giving me a book by Dale Jorgenson with the admonition to "do econometrics like Dale." I hope this dissertation doesn't stray too far from that admonition. Jay Atlas was a constant fountain of sage advice, the content of which I increasingly appreciate with the passage of time.

Finally, this long academic journey would not have been possible without the constant support and inspiration from my parents, Arthur and Heide. They never ceased to listen, even when a conversation about the true purpose of education turned toward bandwidth selection. Through their example they taught me what it means to educate the heart, the hand, and the mind. While the past few years have been heavy on the last, the inspiration they provide has never been stronger.

Contents

Acknowledgements **v**

1 Dynamic Treatment Regimes **1**

1.1 Introduction 2

1.2 Dynamic Causal Effects and Treatment Regimes 6

1.2.1 Basic setup 6

1.2.2 Simple causal estimands 8

1.2.3 Dynamic treatment regimes 9

1.3 Bayesian Inference for Dynamic Treatment Regimes 11

1.3.1 Bayesian perspective on causal effects 11

1.3.2 The science and assignment mechanism 12

1.3.3 Ignorable treatment assignment and sequential unconfoundedness 13

1.3.4 Posterior inference under ignorable treatment assignment . . . 14

1.4 Optimal Treatment and Feasible Outcomes 16

1.5 Implementing the Bayesian Approach 17

1.5.1 Model of the science and priors 17

1.5.2 Simulation-based optimal treatment and feasible outcomes . . 19

1.6 Application: Educational Tracking 21

1.6.1 Data 21

1.6.2 Priors and computation 22

1.6.3 Basic results 23

1.6.4 Principal stratification on intermediate outcomes 24

1.6.5 Ignoring dynamic selection or intermediate effects 26

1.6.6 Optimal treatment and feasible outcome sets 27
1.7 Conclusion 31
1.A Appendix 32
1.A.1 Inference given randomized treatment or no intermediate effects 32
1.A.2 Data and selection 33
1.A.3 Model checking 34
1.A.4 Extended results 36
1.A.5 Complications: peer effects and non-ignorable treatment . . . 40
1.A.6 Posterior inference using MCMC 41

2 Boundary Regression Disconinuity Design 45
2.1 Introduction 46
2.2 The Regression Discontinuity Model 48
2.2.1 Scalar RD designs 48
2.2.2 Boundary RD designs 51
2.3 Estimation 58
2.3.1 Local linear estimation of conditional effects 58
2.3.1.1 Sharp conditional effect 58
2.3.1.2 Fuzzy conditional effect 61
2.3.2 Estimation of average effects 63
2.3.3 Graphical approaches to boundary RD designs 65
2.4 Bandwidth Selection 66
2.5 Application 72
2.5.1 Data 72
2.5.2 Results 73
2.5.2.1 First stage 73
2.5.2.2 Conditional effects 74
2.5.2.3 Average effects 76
2.6 Conclusion 81
2.A Appendix 82
2.A.1 Proofs 82

2.A.2 Nonparametric delta method variance estimator 85
2.A.3 Optimal bandwidth selection 89

3 Do Value-Added Estimates Add Value? 93
3.1 Introduction 94
3.2 Empirical Learning Framework 97
3.2.1 Addressing Child-Level Heterogeneity: Dynamic Panel Approaches to the Education Production Function 100
3.2.2 Addressing Measurement Error in Test Scores 102
3.3 Data 103
3.4 Results 107
3.4.1 Cross-sectional and Graphical Results 107
3.4.1.1 Baseline estimates from cross-section data 107
3.4.1.2 Graphical and reduced-form evidence 109
3.4.2 OLS and Dynamic Panel Value-Added Estimates 115
3.4.2.1 The persistence parameter 115
3.4.2.2 The contribution of private schools 116
3.4.2.3 Regression diagnostics 117
3.5 Conclusion 118
3.A Additional Estimation Strategies 121
3.A.1 System GMM 121
3.A.1.1 Uncorrelated or constantly correlated effects 121
3.A.1.2 Conditional mean stationarity 122
3.A.1.3 Results 123
3.A.2 Attrition corrected estimators 124

Bibliography 126

Chapter 1

Bayesian Inference for Dynamic Treatment Regimes: Mobility, Equity, and Efficiency in Student Tracking [1]

Policies in health, education, and economics often unfold sequentially and adapt to changing conditions. Such time-varying treatments pose problems for standard program evaluation methods because intermediate outcomes are simultaneously pre-treatment confounders and post-treatment outcomes. This paper extends the Bayesian perspective on causal inference and optimal treatment to these types of dynamic treatment regimes. A unifying idea remains ignorable treatment assignment, which now sequentially includes selection on intermediate outcomes. I present methods to estimate the causal effect of arbitrary regimes, recover the optimal regime, and characterize the set of feasible outcomes under different regimes. I demonstrate these methods through an application to optimal student tracking in ninth and tenth grade mathematics. For the sample considered, student mobility under the status-quo regime is significantly below the optimal rate and existing policies reinforce between student inequality. An easy to implement optimal dynamic tracking regime, which promotes more students to honors in tenth grade, increases average final achievement 0.07 standard deviations above the status quo while lowering inequality; there is no binding equity-efficiency tradeoff. The proposed methods provide a flexible

[1]This chapter has benefited from comments by Alberto Abadie, Guido Imbens, Dale Jorgenson, Asim Khwaja, Jamie Robins, Donald Rubin, and seminar participants at Harvard. Thanks go to the NCERDC for providing data, and IQSS and HMDC for providing computing resources.

and principled approach to causal inference for time-varying treatments and optimal treatment choice under uncertainty.

1.1 Introduction

Programs in health, education, and economics can often be described as dynamic treatment regimes—adaptive policies that recommend actions in each treatment period depending on past observations and decisions. Economists, for instance, recommend dynamic social insurance programs that account for changing needs and incentives (e.g., Shavell and Weiss, 1979; Heckman et al., 1999; Fredriksson and Holmlund, 2006). Educators seek instructional regimes, implemented over many years, and tailored to each child's development, that best advance educational goals (e.g., Hong and Raudenbush, 2008; Raudenbush, 2008). And epidemiologists debate how to sequence and assign treatment to individual patients (e.g., Murphy, 2003; Robins, 2004; Kitahata et al., 2009; Lang, 2009). Policymakers, across many disciplines, require estimates of the causal effect of such policies and a means to select between them. For instance, how should unemployment benefits be scheduled to maximize social welfare? Or how should schools track students to maximize average performance? Or when should patients begin HIV antiretroviral or Parkinson's treatment given their developing health status?

As highlighted by Robins (1986), time-varying treatments pose problems for standard program evaluation methods because intermediate variables are simultaneously post-treatment outcomes and pre-treatment confounders. An HIV-positive patient's measured health status, for example, depends on previous treatments and influences future treatment. Likewise, a student's test score depends on past coursework and influences future coursework. Given time-varying treatments, the standard advice to control for pre-treatment covariates but not post-treatment outcomes no longer makes sense. Not controlling for intermediate outcomes ignores an important determinant of selection into treatment, but controlling for intermediate outcomes ignores that

they are affected by earlier treatments. Analysis of time-varying treatments, and dynamic treatment regimes more generally, therefore require methodological tools that can properly incorporate intermediate outcomes and dynamic selection.

This paper presents the Bayesian perspective on causal inference for time-varying treatments and dynamic treatment regimes. Once properly formulated, dynamic treatment regimes fit within the unifying Bayesian framework for causal inference introduced by Rubin (1978). This approach defines causal effects in terms of potential outcomes, clearly separates the scientific model for potential outcomes and covariates from the treatment assignment mechanism, and then, optionally, uses Bayesian posterior predictive inference on causal effects (Rubin, 1978, 2008). Sequential selection into treatment on intermediate outcomes can be addressed in this framework.

A unifying idea remains that all ignorable treatment mechanisms—mechanisms that are independent of missing potential outcomes but may depend on the observed data—yield the same posterior inferences and that sequential unconfoundedness is a particular ignorable mechanism. This implies that existing multiple imputation methods can be used *after* carefully defining potential outcomes and the observed and missing data provided that sequential unconfoundedness holds. I present a model that can account for discrete and continuous variables with grouped structures. The partial pooling of information across groups offered by Bayesian methods plays an important role in providing information from which to impute missing potential outcomes. Using this model, I describe estimation of simple causal estimands, such as the comparison between two treatment sequences, more complex estimands, such as the comparison between two dynamic treatment regimes, and optimal treatment choice, such as when to begin treatment based on time-varying characteristics.

I approach optimal treatment choice as a Bayesian decision problem (Berger, 1985; Dehejia, 2005). Policymakers seek the optimal treatment rule for future units that have not yet been assigned treatment. The optimal treatment rule can depend both on units' baseline characteristics and time-varying intermediate outcomes. Consistent with practical policy constraints, the rules may be limited to a feasible class, such as those based on a simple linear index or a single covariate. By adopting a Bayesian

perspective, the resulting regime is optimal regardless of sample size and integrates over all sources of uncertainty.

I demonstrate the proposed approach using ninth and tenth grade mathematics tracking in North Carolina public schools. Tracking policies can have significant impacts on average student performance and between-student inequality and remain controversial. Common concerns include tracking students into inappropriate courses, tradeoffs between equity and efficiency of outcomes, and insufficient mobility between tracks (Slavin, 1987, 1990; Hanushek and Woessmann, 2005; Brunello and Checchi, 2007; Duflo et al., 2010). As in most other school systems, students in North Carolina can choose different levels of specific courses, such as standard and honors Algebra I, Geometry, and Algebra II. Performance in Algebra I strongly influences track assignments in Geometry and Algebra II.

Using data on students who enroll in Algebra I in ninth grade, I find significant tracking sequence effects. Amongst single-streaming policies, enrolling all students to standard mathematics in ninth grade and honors mathematics in tenth grade yields the highest average achievement, roughly 0.13 standard deviations higher than demotion from honors to a standard and 0.07 standard deviations above the status quo. Methods that ignore dynamic selection or intermediate causal effects yield significantly different results.

To study treatment choice, I consider three classes of optimal treatment rules: a dynamic cutoff rule that assigns students to honors if their most recent score exceeds a threshold, a static index rule that assigns students to honors based on their baseline characteristics, and a dynamic index rule that assigns students based on all past observed data. All three rules increase average achievement while lowering inequality relative to the status quo. Most of these gains arise from streaming students into the standard/honors tracking sequence. By comparison, the status quo policy significantly under-promotes students after ninth grade, assigning over 60 percent to the standard/standard track.

The proposed methods can also characterize the set of feasible outcomes. I explore the tradeoff between equity and efficiency, as measured by the mean and standard

deviation of performance obtainable by a dynamic cutoff regime. There is significant room for improvement over the status quo in both directions. Holding average achievement constant, the status quo tracking rule outcome is near the upper feasible bound for achievement variance. These results are consistent with concerns that status quo policies reinforce between-student inequality. But there is no binding equity-efficiency tradeoff.

The approach developed in this paper differs from existing proposals such as g-computation and structural nested mean models pioneered by Robins (1986, 1989, 1997, 1999a) and Gill and Robins (2001), and inverse probability weighting and marginal structural models proposed by Robins (1999b), Robins et al. (2000), Hernan et al. (2001), Murphy et al. (2001), and Lechner (2009). Robins and Hernan (2008) provide an accessible review of g-computation, structural nested mean models, and inverse probability weighting for time-varying exposures. Abbring and Heckman (2007) and Heckman and Navarro (2007) link this literature to economic dynamic discrete choice models. This paper's Bayesian approach to optimal treatment choice also differs from plug-in proposals by Murphy (2003) and Robins (2004).

The proposed approach has both practical and theoretical appeal. First, once properly framed, it offers a simple recipe for studying time-varying treatments that enforces a clear separation between the scientific model and treatment assignment mechanism. For many applications, existing software for multiple imputation can be adapted to study time-varying treatments. Second, Bayesian hierarchical models allow partial pooling of information where scientifically justified, both across groups and across treatment sequences, thereby informing the imputation of the missing data on the basis of the pooled observed data. A major concern in sequential settings is the exponential explosion of potential outcomes associated with each treatment sequence. As the number of possible treatment sequences grows, pooling information becomes critical; hierarchical models on the space of potential outcomes provide a flexible alternative that falls between the two extremes—complete pooling or separation—offered by marginal structural models. For applications in education, where grouped structures are common, pooling information across groups is also helpful. Finally, Bayesian methods are ideally suited to study optimal treatment choice. The methods

I propose integrate over all sources of uncertainty—a potentially vital step in small samples or high-dimensional settings where uncertainty is large. While there remain significant opportunities to build richer models and hierarchical priors suitable for larger-scale applications, the proposed approach provides a framework to study time-varying treatment problems that lie at the heart of epidemiology, education, and economics.

1.2 Dynamic Causal Effects and Treatment Regimes

1.2.1 Basic setup

Time-varying treatment applications often include many periods and treatments. However, all the conceptual issues surrounding dynamic treatment regimes can be captured in a simple, two-period, binary treatment model. I focus on this simplified setup because extensions to additional periods and categorical treatments are straightforward but add notational complexity.

Consider a sample of N units, such as children, patients, or workers, indexed by $i = 1, \ldots, N$. In each of two periods, indexed by $t = 1, 2$, units receive a binary treatment W_{it}. By convention, let $W_{it} = 1$ denote receiving the active treatment and $W_{it} = 0$ denote receiving the control or placebo treatment. Units can therefore experience treatment in both periods $(1, 1)$, neither period $(0, 0)$, only the first period $(1, 0)$, or only the second period $(0, 1)$. For instance, these four sequences could represent the standard and honors mathematics tracks in ninth and tenth grade. Alternatively, if $(1, 0)$ is not an available option, the three remaining sequences could represent varying treatment start times for Algebra I coursework, or, applied to health, initiation of HIV or Parkinson's therapy. Treatment sequences are a type of compound treatment where the constituent treatments unfold sequentially.

Units have baseline ($t = 1$) covariates and intermediate ($t = 2$) outcomes X_{it} drawn from the space $\mathcal{X}_t$. Baseline covariates include characteristics measured prior to the onset of treatment, such as age, gender, baseline test scores, or an initial health measure. Intermediate outcomes, such as an updated achievement or health measure,

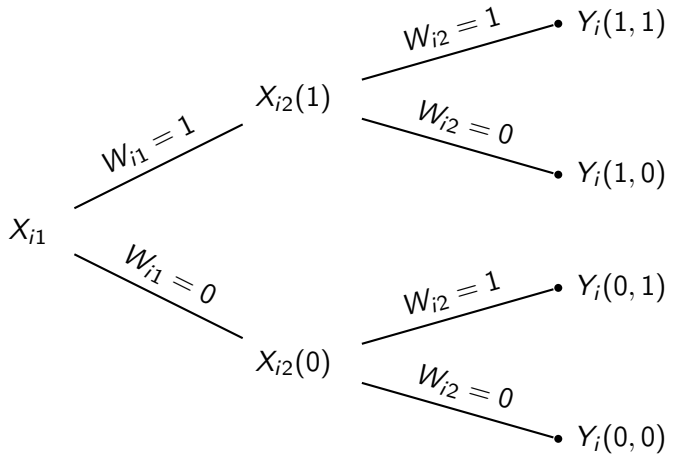


Figure 1.1: A two-period, time-varying, binary treatment setup. Each path along the tree represents a potential treatment sequence. Additional periods or treatments grow the tree structure.

are recorded after the first treatment but prior to the second treatment. Thus X_{i2} comes before W_{i2} even though both have $t = 2$. A final outcome Y_i, such as tenth grade achievement, is the object of primary interest and is measured after the final treatment. Covariates and outcomes may be multivariate, although I reserve boldface to denote vectors and matrices of these primitives.

Following Neyman (1923), Rubin (1974, 1978), and Robins (1986), I define causal effects in terms of potential outcomes. Potential outcomes capture the hypothetical outcomes associated with a particular, not necessarily assigned, treatment sequence. With treatment unfolding over multiple periods, both intermediate outcomes X_{i2} and final outcomes Y_i have associated potential outcomes. Using lower case letters to denote fixed values, $X_{i2}(w_1)$ and $Y_i(w_1, w_2)$ represent the intermediate and final potential outcomes under the hypothetical treatment sequence (w_1, w_2). For instance, $X_{i2}(1)$ gives unit i's intermediate outcome assuming treatment in the first period, and $Y_i(1, 1)$ gives the final outcome assuming treatment in both periods. Implicit in this multi-period potential outcome notation is the assumption that treatments cannot have effects before they occur and there is no interaction between units (i.e., SUTVA in Rubin, 1980). Figure 1.1 illustrates this two-period, binary treatment setup.

A key feature of potential outcomes is that we observe only the potential outcomes associated with the treatment sequence units actually receive. Thus *observed* outcomes link to *potential* outcomes as

$$X_{i2} \equiv X_{i2}(W_{i1}), \qquad Y_i \equiv Y_i(W_{i1}, W_{i2}). \tag{1.1}$$

For every observed potential outcome there are one or more *missing* potential outcomes. Specifically, the missing intermediate potential outcomes are

$$X_{i2}^{\text{mis}} \equiv X_{i2}(1 - W_{i1}), \tag{1.2}$$

and the missing final outcomes are

$$\mathbf{Y}_i^{\text{mis}} \equiv (Y_i(1 - W_{i1}, 1 - W_{i2}), Y_i(W_{i1}, 1 - W_{i2}), Y_i(1 - W_{i1}, W_{i2}))\,. \tag{1.3}$$

Equations (1.1) through (1.3) transform potential outcomes and observed treatment to observed and missing data. In any particular sample, the observed treatment assignment indicators uniquely determine the missing data matrix that indicates which values are missing and observed.

Compared to point treatments, there are substantially more missing potential outcomes than observed outcomes. However, the primary difficulty remains the same: since units can only receive one treatment sequence at any time, only one sequence of potential outcomes is ever observed. The inability to observe all potential outcomes defines the "fundamental problem of causal inference" (Holland, 1986) and leads to a view of causal inference as a missing data problem (Rubin, 1978).

1.2.2 Simple causal estimands

Causal effects are comparisons between potential outcomes for a common set of units (Rubin, 2005). For ease of exposition, I focus on the final outcome $Y_i(w_1, w_2)$ and average effects. However, we can easily define other estimands based on intermediate outcomes or ratios and quantiles.

At the lowest level, comparisons between $Y_i(w_1, w_2)$ and $Y_i(w_1', w_2')$ give the *unit-level causal effect* of treatment (w_1, w_2) compared to treatment (w_1', w_2'). In general, causal effects are defined over collections of units. Following the literature on point-treatment causal effects, let the *sample-average treatment effect* be

$$\tau_{\text{SATE}}(\mathbf{w}, \mathbf{w}') \equiv \frac{1}{N} \sum_{i=1}^{N} \left[Y_i(w_1, w_2) - Y_i(w_1', w_2') \right]. \tag{1.4}$$

I focus primarily on inference for sample average effects rather than population inference. Alternative causal estimands, such as quantile treatment effects or conditional treatment effects, are simply different comparisons between potential outcomes or comparisons within different subsamples.

1.2.3 Dynamic treatment regimes

Although treatments index potential outcomes, it is also useful to think of potential outcomes in terms of treatment regimes. A treatment regime is an assignment mechanism that determines treatment for each unit on the basis of previous observations for that unit. The biostatistics literature also refers to treatment regimes as adaptive strategies, interventions, treatments, or therapies.

I define a *dynamic treatment regime* as a pair $\boldsymbol{\delta}$ of decision functions $\delta_1 : \mathcal{X}_1 \mapsto \{0, 1\}$ and $\delta_2 : \mathcal{X}_2 \times \{0, 1\} \times \mathcal{X}_1 \mapsto \{0, 1\}$ that assign units with observed covariates (x_2, w_1, x_1) to a treatment sequence $\boldsymbol{\delta} \equiv (\delta_1(x_1), \delta_2(x_2, w_1, x_1))$. More generally, Murphy et al. (2001) study randomized dynamic treatment regimes where decision functions are conditional probability distributions over treatments. Effects of randomized treatment regimes are straightforward to estimate in a Bayesian framework but, again, they add notational complexity and are rarely of interest to policymakers.

We can distinguish between *dynamic* treatment regimes that assign treatment based on time-varying covariates, X_{i1} and X_{i2}, and *static* treatment regimes that assign treatment based only on baseline covariates, X_{i1}. As an example, assigning

ninth graders to an honors track if they score above a cutoff on a baseline exam and tenth graders to honors if they receive an A or B in ninth grade is a dynamic treatment regime, whereas assigning students based only on their baseline socio-economic status or achievement is a static treatment regime. In some cases, we may be interested in treatment regimes of a specific class $\mathcal{D}$. Robins et al. (2008) and Lok et al. (2008) study regimes that initiate treatment if a single time-varying covariate falls below of pre-specified threshold. Restricted rules may be of interest for legal or normative reasons, such as to avoid racial profiling, or to simplify implementation.

Because treatment regimes map covariate histories to treatments, we can index potential outcomes by decision functions instead of treatment sequences. That is, $X_{i2}\left(\delta_1(X_{i1})\right)$ gives the intermediate potential outcome given treatment assignment $\delta_1(X_{i1})$ and

$$Y_i(\boldsymbol{\delta}) \equiv Y_i(\delta_1(X_{i1}), \delta_2(X_{i2}(\delta_1(X_{i1})), \delta_1(X_{i1}), X_{i1}))$$

gives final potential outcome under regime $\boldsymbol{\delta}$.

To define the average causal effects between two regimes, let $\boldsymbol{\delta}'$ be the reference regime such as a placebo treatment in all periods or never receiving treatment. Then the *sample-average treatment regime effect* is

$$\tau_{\text{SATRE}}(\boldsymbol{\delta}, \boldsymbol{\delta}') \equiv \frac{1}{N} \sum_{i=1}^{N} \left[Y_i(\boldsymbol{\delta}) - Y_i(\boldsymbol{\delta}') \right]. \tag{1.5}$$

If $\boldsymbol{\delta}$ is a particular static regime that treats all units equally, say $\boldsymbol{\delta} = (w_1, w_2)$, and the reference regime assigns $\boldsymbol{\delta}' = (w_1', w_2')$, then the average treatment regime effect is equivalent to the average treatment effect. Again, we can easily define treatment regime effects in terms of ratios, quantiles, or any other comparison between potential outcomes.

In applications with existing status quo policies, such as existing mathematics tracking rules, we may also be interested in the causal effect of a new policy compared

to the status quo regime. For instance, does the status quo tracking policy perform better than single-streaming children into a common track? To capture this idea, we can define the *sample-average treatment regime improvement* as

$$\tau_{\text{SATRI}}(\boldsymbol{\delta}) \equiv \frac{1}{N} \sum_{i=1}^{N} \left[Y_i(\boldsymbol{\delta}) - Y_i \right], \tag{1.6}$$

where we define improvement over the observed status quo outcome Y_i. Likewise, if we are interested in the causal effect of regimes on inequality, measured by the sample variance, we can define the *sample-variance treatment regime improvement* as

$$\tau_{\text{SVTRI}}(\boldsymbol{\delta}) \equiv \frac{1}{N} \sum_{i=1}^{N} \left[Y_i(\boldsymbol{\delta}) - \frac{1}{N} \sum_{i=1}^{N} Y_i(\boldsymbol{\delta}) \right]^2 - \frac{1}{N} \sum_{i=1}^{N} \left[Y_i - \frac{1}{N} \sum_{i=1}^{N} Y_i \right]^2, \tag{1.7}$$

where negative values, i.e., lower sample variance, indicates an improvement.

1.3 Bayesian Inference for Dynamic Treatment Regimes

1.3.1 Bayesian perspective on causal effects

Rubin (1978) introduced the Bayesian perspective on causal inference and its relation to general missing data problems (Rubin, 1976). Rubin (2005, 2008), Imbens and Rubin (1997, forthcoming), and Hirano et al. (2000) provide further discussion and extensions of causal inference from a Bayesian perspective. To a Bayesian, the missing potential outcomes are no different than unknown parameters. The natural solution is therefore to form a posterior predictive distribution over missing outcomes.

Despite the conceptual appeal and inherent simplicity of the Bayesian perspective, its application has not been fully articulated for time-varying treatments.

Bayesian inference for sample causal effects on the final and intermediate outcomes $\mathbf{Y}$ and $\mathbf{X}_2$, where the missing i subscript denotes the full data, i.e., $\mathbf{Y} \equiv (Y_1, Y_2, \ldots, Y_N)$, follows from the posterior predictive distribution of missing potential outcomes $f(\mathbf{Y}^{\text{mis}}, \mathbf{X}_2^{\text{mis}} \mid \mathbf{Y}, \mathbf{W}_2, \mathbf{X}_2, \mathbf{W}_1, \mathbf{X}_1)$. The posterior predictive distribution allows us to "fill in" or multiply-impute the missing data. Typically we evaluate the predictive distribution and resulting inference on causal effects by simulation. Focusing on the final outcomes, let $Y_i^{\text{mis},(l)}(w_1, w_2)$ denote a draw l from the posterior predictive distribution of the missing outcomes under treatment (w_1, w_2). A "completed" draw l for unit i is then

$$Y_i^{(l)}(w_1, w_2) \equiv \begin{cases} Y_i & \text{if } W_{i1} = w_1 \text{ and } W_{i2} = w_2, \\ Y_i^{\text{mis},(l)}(w_1, w_2) & \text{otherwise.} \end{cases}$$

Posterior distributions for sample causal estimands are functions of the completed data.

Treatment regime effects can be estimated via imputation just as easily as static causal effects. For sample-average treatment regime effects, τ_{SATRE}, we can draw from the posterior predictive distribution to complete the data, and then select the relevant potential outcomes for any treatment rule $\boldsymbol{\delta}$ as $Y_i^{(l)}(\boldsymbol{\delta}) \equiv Y_i^{(l)}(\delta_2(X_{i2}^{(l)}(\delta_1(X_{i1})), \delta_1(X_{i1}), X_{i1}), \delta_1(X_{i1}))$.

1.3.2 The science and assignment mechanism

Obtaining the posterior predictive distribution requires a model of the data. A basic insight of Rubin (1978) is the joint distribution of potential outcomes, assignments, and covariates factors into two components, which Rubin calls the science and the treatment assignment mechanism. In Rubin's terminology, the science represents the true underlying data, whereas the assignment mechanism determines what data are

actually observed or missing. The same general principle applies to time-varying treatment. Let $\mathbf{S} \equiv (\mathbf{Y}(0,0), \mathbf{Y}(0,1), \mathbf{Y}(1,0), \mathbf{Y}(1,1), \mathbf{X}_2(0), \mathbf{X}_2(1), \mathbf{X}_1)$ denote the underlying data of interest. In the two-period, dynamic-treatment setup, we can factor the joint distribution of missing and observed data, denoted $\mathbf{J}$, into

$$f(\mathbf{J}) = f(\mathbf{S})\, f(\mathbf{W}_2, \mathbf{W}_1 \mid \mathbf{S}). \tag{1.8}$$

This factorization clearly separates the subject matter model or science from the particular treatment assignment mechanism that determines what we are able to observe.

1.3.3 Ignorable treatment assignment and sequential unconfoundedness

In order to make progress on causal inference, we must make assumptions about the treatment assignment mechanism. Rubin (1976, 1978) defines a missing data mechanism, such as treatment assignment, as *ignorable* if it does not depend on the missing data. Adapted to time-varying treatments, the weakest form of ignorable treatment is

$$f(\mathbf{W}_2, \mathbf{W}_1 \mid \mathbf{S}) = f(\mathbf{W}_2, \mathbf{W}_1 \mid \mathbf{Y}, \mathbf{X}_2, \mathbf{X}_1). \tag{1.9}$$

That is, treatment assignments can depend on all the recorded data $(\mathbf{Y}, \mathbf{X}_2, \mathbf{X}_1)$ but not the missing data $(\mathbf{X}_2^{\text{mis}}, \mathbf{Y}^{\text{mis}})$. Completely randomized experiments and experiments randomized conditional on baseline characteristics $\mathbf{X}_1$ are clearly ignorable. But they are both special cases of ignorability that fail in most time-varying applications due to dynamic confounding.

Much of the literature on time-varying treatments focuses on a particular ignorable assignment mechanism: sequential unconfoundedness. Formally, Robins (1986) defines *sequential unconfoundedness* as

$$f(\mathbf{W}_2, \mathbf{W}_1 \mid \mathbf{S}) = f(\mathbf{W}_2 \mid \mathbf{X}_2, \mathbf{W}_1, \mathbf{X}_1) f(\mathbf{W}_1 \mid \mathbf{X}_1). \tag{1.10}$$

That is, observed treatments $\mathbf{W}_1$ and $\mathbf{W}_2$ are independent of missing past potential outcomes and future potential outcomes given past observed data. Sequential unconfoundedness is intuitive in longitudinal settings and is substantially weaker than assuming that the entire treatment sequence is independent of the potential outcomes conditional on baseline characteristics. For instance, it fits teachers that assign students randomly to tracks conditional on previous tracking and observed performance, or doctors that propose therapies randomly conditional on observed prognostic factors and prior treatments up to that point. Critically, sequential unconfoundedness is an ignorable assignment mechanism because it does not depend on missing data.

1.3.4 Posterior inference under ignorable treatment assignment

Rubin (1978, 2008) describes posterior inference in point-treatment settings under ignorable treatment assignment. Once set up in a similar framework, posterior inference for time-varying treatments follows the same basic steps. As in point-treatment settings, posterior inference for sample causal effect follows from the posterior predictive distribution over missing data,

$$f(\mathbf{Y}^{\text{mis}}, \mathbf{X}_2^{\text{mis}} \mid \mathbf{Y}, \mathbf{W}_2, \mathbf{X}_2, \mathbf{W}_1, \mathbf{X}_1) = \frac{f(\mathbf{S})\, f(\mathbf{W}_2, \mathbf{W}_1 \mid \mathbf{S})}{\int\int f(\mathbf{S})\, f(\mathbf{W}_2, \mathbf{W}_1 \mid \mathbf{S})\, d\mathbf{Y}^{\text{mis}} d\mathbf{X}_2^{\text{mis}}}. \quad (1.11)$$

Under any ignorable assignment mechanism, including sequential unconfoundedness, (1.11) simplifies to

$$f(\mathbf{Y}^{\text{mis}}, \mathbf{X}_2^{\text{mis}} \mid \mathbf{Y}, \mathbf{W}_2, \mathbf{X}_2, \mathbf{W}_1, \mathbf{X}_1) = \frac{f(\mathbf{S})}{\int\int f(\mathbf{S})\, d\mathbf{Y}^{\text{mis}} d\mathbf{X}_2^{\text{mis}}} \propto f(\mathbf{S}). \quad (1.12)$$

Ignorable treatment rules, where treatment is independent of the missing potential outcomes, allow the treatment rule to move from under the integral and cancel with the treatment rule in the numerator.

This result illustrates the unifying scope of ignorable missing data mechanisms. Causal inference for dynamic treatment regimes under ignorable treatment assignment, including sequential unconfoundedness, requires only a model of the science $f(\mathbf{S})$. After appropriately defining the potential outcomes for intermediate and final outcomes, and therefore the missing and observed data, a full-information analysis proceeds equivalently to standard multiple imputation procedures under a missing-at-random/ignorability assumption (e.g., Little and Rubin, 1987). A major benefit of this result is that existing software for multiple imputation can often be used, even if not originally designed with sequential unconfoundedness in mind.

So far the model has no explicit parameters and instead works with the full data. We can make the model practical by appealing to exchangability of the unit indices. Assuming that the parameters and priors for the science and assignment mechanism are distinct, i.e., $\boldsymbol{\theta}_J \equiv (\boldsymbol{\theta}, \boldsymbol{\theta}_W)$ and $\pi(\boldsymbol{\theta}, \boldsymbol{\theta}_W) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_W)$, the the logic of (1.11) and (1.12) still applies and we can write the model of science as

$$f\left(\mathbf{S}\right) = \int \left[\prod_{i=1}^{N} f(Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), X_{i2}(0), X_{i2}(1), X_{i1} \mid \boldsymbol{\theta}) \right] \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

The applied problem, which I confront in Section 1.5, is how best to model $f(\mathbf{S}_i \mid \boldsymbol{\theta})$, specify the prior $\pi(\boldsymbol{\theta})$, and compute the posterior distribution for parameters and missing data. This problem may present significant practical difficulties, particularly in applications with many periods and treatments, but is conceptually straightforward.

1.4 Optimal Treatment and Feasible Outcomes

Policymakers, in health, education, or economics, primarily seek the optimal treatment regime not simply the effect of an arbitrary regime. Manski (2000), Manski (2004), Dehejia (2005), and Hirano and Porter (2009) study frequentist and Bayesian decision theoretic approaches to optimal treatment in the context of point treatments. For dynamic treatment regimes, Murphy (2003), Robins (2004), Robins et al. (2008), and Lok et al. (2008) propose semiparametric plug-in methods to find the regime that maximizes the expected final outcome $\mathbb{E}\left[Y_i(\boldsymbol{\delta})\right]$ asymptotically.

I consider a Bayesian decision theoretic approach to optimal treatment choice. Assume data analysts observe a random sample of N units with data $\mathbf{Z} \equiv (\mathbf{Y}, \mathbf{W}_2, \mathbf{X}_2, \mathbf{W}_1, \mathbf{X}_1)$ from which they infer treatment efficacy and base recommendations for future units. Constraints limit policymakers to a class of feasible rules $\mathcal{D}$, and policymakers have a utility function over outcomes.

Information about a treatment regime's efficacy is contained in the posterior predictive distribution $f(\tilde{Y}(\boldsymbol{\delta}) \mid \mathbf{Z})$, where the tilde symbolizes a hypothetical future unit and $\boldsymbol{\delta}$ is a function imputed data. If we have preferences $u_{\boldsymbol{\delta}}(y)$ over the final outcome, subscripted by $\boldsymbol{\delta}$ to allow for treatment costs, then outcome of interest – the *posterior expected utility* – for regime $\boldsymbol{\delta}$ given prior $\pi(\boldsymbol{\theta})$ is

$$U(\boldsymbol{\delta}, \pi \mid \mathbf{Z}) \propto \int \int u_{\boldsymbol{\delta}}\left(\tilde{Y}\left(\boldsymbol{\delta}\right)\right) f(\tilde{Y}\left(\boldsymbol{\delta}\right) \mid \boldsymbol{\theta}) \prod_{i=1}^{N} \left[f(\mathbf{S}_i \mid \boldsymbol{\theta})\right] \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} d\tilde{Y}\left(\boldsymbol{\delta}\right). \tag{1.13}$$

Expected utility averages over the posterior distribution of the unknown outcome $\tilde{Y}\left(\boldsymbol{\delta}\right)$ conditional on the observed data $\mathbf{Z}$ and incorporate uncertainty in the parameters $\boldsymbol{\theta}$.

The optimal treatment regime selects the maximizing rule, conditional on the observed data and prior, $\boldsymbol{\delta}^*(\pi, \mathbf{Z}) = \arg\max_{\boldsymbol{\delta} \in \mathcal{D}} U(\boldsymbol{\delta}, \pi \mid \mathbf{Z})$. How best to compute $\boldsymbol{\delta}^*$ depends on the class of feasible rules $\mathcal{D}$. But the general strategy is always the same:

integrate over all sources of uncertainty by maximizing using the posterior predictive distribution.

Given an estimated optimal policy rule $\boldsymbol{\delta}^*$, we may be interested in the outcome under that rule. The population-average treatment regime effect can be estimated by applying the optimal rule to the posterior predictive simulations and comparing it to another benchmark regime. Alternatively, we could apply the optimal rule retrospectively to the sample data and ask, for example, whether the optimal treatment rule performs better than the status quo policy on the observed sample. To capture this idea, define the *sample-average maximum improvement* as $\tau_{\text{SAMI}} = N^{-1} \sum_{i=1}^{N} \left[Y_i(\boldsymbol{\delta}^*) - Y_i \right]$.

In some instances, we may wish to characterize the set of feasible outcomes that are (ex-ante) achievable given a particular class of rules. Consider the perceived tradeoff between equity and efficiency in education. Some educators worry that tracking rules designed to maximize average achievement increase inequality. The set of (ex-ante) feasible mean and variance outcomes is, $\left\{ \mathbb{E}\left[\tilde{Y}(\boldsymbol{\delta}) \mid \mathbf{Z} \right], \mathbb{V}\left[\tilde{Y}(\boldsymbol{\delta}) \mid \mathbf{Z} \right] : \boldsymbol{\delta} \in \mathcal{D} \right\}$, where the conditional expectation and variance is taken over the posterior predictive distribution given observed data $\mathbf{Z}$ and prior π. To characterize the boundary of this set, we can define a mean-variance utility function as $U(\boldsymbol{\delta}, \pi \mid \boldsymbol{\lambda}; \mathbf{Z}) = \lambda_1 \cdot \mathbb{E}\left[\tilde{Y}(\boldsymbol{\delta}) \mid \mathbf{Z} \right] + \lambda_2 \cdot \mathbb{V}\left[\tilde{Y}(\boldsymbol{\delta}) \mid \mathbf{Z} \right]$, with weights λ_1 and λ_2 on the mean and variance terms capturing different preferences. By varying λ_1 and λ_2 and maximizing we can recover the optimal rule for any mean-variance preference and therefore the boundary of feasible mean and variance outcomes.

1.5 Implementing the Bayesian Approach

1.5.1 Model of the science and priors

The most appropriate parametric model of $f(\mathbf{S}_i \mid \boldsymbol{\theta})$ and prior $\pi(\boldsymbol{\theta})$ depends on the precise application. For the application to education tracking, I consider a model directly on joint distribution rather than an alternative factorization. Directly modeling the joint distribution leads to a simple and efficient sampling strategy. An

alternative would be to model the data conditional on baseline covariates or to use chained-equation approaches commonly found in the multiple imputation literature (e.g., Su et al., 2009).

The education data I study includes both continuous variables (e.g., test scores and age) and discrete variables (e.g., gender and free lunch eligibility). Discrete data can be modeled using a latent variable formulation. Let $\mathbf{S}_i^*$ represent both the continuous data and a latent representation of the discrete data. That is, $\mathbf{S}_i^* = (\mathbf{S}_i^c, \mathbf{S}_i^{d*})$ where $\mathbf{S}_i^{d*}$ are latent variables underlying the discrete components $\mathbf{S}_i^d$, and $\mathbf{S}_i^c$ are the continuous variables. If the discrete variables are binary, i.e., $\mathbf{S}_{ik}^d \in \{0, 1\}$ for each binary component k, then we can link the observed discrete data to their latent variables through $\mathbf{S}_{ik}^d = \mathbf{1}\{S_{ik}^{d*} \geq 0\}$. Ordinal data can be modeled with additional cutoffs.

A second common characteristic of education data is a nested structure. Students reside within schools and classrooms. To account for school-level heterogeneity, I introduce a random-effects multivariate normal model:

$$\text{within school } j\text{: } \mathbf{S}_i^* \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{1.14}$$

where $\boldsymbol{\mu}_j$ is a $K \times 1$ vector of means, $\boldsymbol{\Sigma}_j$ is a $K \times K$ covariance matrix that varies by school j, and K is the dimension of $\boldsymbol{S}_i$. I place conditionally conjugate hierarchical priors on the school means and covariances, $\boldsymbol{\mu}_j \sim \mathcal{N}(\mathbf{m}, \mathbf{V})$ and $\boldsymbol{\Sigma}_j \sim \mathcal{IW}(\nu, \mathbf{T})$, where $\mathbf{m}$ and $\mathbf{V}$ are the mean and covariance matrix for a multivariate-Normal distribution, and ν and $\mathbf{T}$ are the degrees of freedom and scale matrix for an inverse-Wishart distribution. Given the internal replication in the data generated by schools, these parameters can be estimated from the data in a hierarchical setup. To complete the model, I place hyperpriors $\mathbf{m} \sim \mathcal{N}(\mathbf{a}_m, \mathbf{B}_M)$, $\mathbf{V} \sim \mathcal{IW}(a_V, \mathbf{B}_V)$, $\nu - K - 1 \sim \ln\mathcal{N}(a_\nu, b_\nu)$, and $\mathbf{T} \sim \mathcal{W}(a_T, \mathbf{B}_T)$ at the top level.

Both the intercept and slopes of the conditional distributions for missing potential outcomes, which are central to imputing missing data, vary by school and treatment sequence but are shared within schools. In this setup, the central untestable assumption is that sequential unconfoundedness holds within schools. With so many

parameters, the hierarchical priors play the important role of pooling information across schools. Intuitively, the degrees of freedom parameter ν captures the numbers of prior measurements – or pseudo-observations – on $\mathbf{\Sigma}_j$ and determines the degree of pooling. In applications with many time periods or grouped data, the ability to use hierarchical priors that partially pool information across similar treatment paths or groups is a potential major advantage of Bayesian methods. Existing strategies, such as marginal structural models, enforce complete pooling or separation of potential outcomes and do not easily accommodate grouped structures.

Estimation of the random-effects multivariate normal model follows standard Markov Chain Monte Carlo (MCMC) methodology. The basic idea is to vary between data augmentation steps, which completes the data, and parameter updates, which are straightforward given the completed data. Given lower level parameters, the hierarchical priors are then updated either using a Gibbs step or a slice-sampling step, depending on the ability to draw from the full conditionals. An appendix contains a full description of the MCMC algorithm and posterior predictive model checks that do not suggest any problems with the model specification.

1.5.2 Simulation-based optimal treatment and feasible outcomes

A simple simulation strategy can be used to estimate the optimal treatment rule given a parametric class of policy rules $\mathcal{D}$. Let β_1 be parameters for the first period rule $\delta_1(x_1;\beta_1)$ and β_2 be the parameters for the second period rule $\delta_2(x_2, w_1, x_1;\beta_2)$. For instance, δ_1 and δ_2 may assign treatment if a linear index $x_1'\beta_1$ and $(x_2', \delta_1(x_1;\beta_1), x_1')'\beta_2$ exceeds a cutoff. In Section 1.6, I consider static and dynamic linear index rules and a dynamic cutoff rule. However, the following strategy can be used for any parametric rule, regardless of flexibility.

The estimation strategy consists of two steps. First, we simulate from the joint posterior predictive distribution,

$$\int f(\tilde{Y}(0,0), \tilde{Y}(0,1), \tilde{Y}(1,0), \tilde{Y}(1,1), \tilde{X}_2(0), \tilde{X}_2(1), \tilde{X}_1 \mid \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid \mathbf{Z}) d\boldsymbol{\theta},$$

for a large number R of hypothetical future units, which we index with $i > N$. Given this approximation to the population predictive distribution, we search for the optimal rule

$$\hat{\boldsymbol{\delta}}^*(\hat{\beta}_1, \hat{\beta}_2) = \arg\max_{\boldsymbol{\delta}(\beta_1,\beta_2)} \frac{1}{R} \sum_{i=N+1}^{N+R} u_{\boldsymbol{\delta}} \left(\tilde{Y}_i \left(\delta_1(\tilde{X}_{i1}; \beta_1), \delta_2(\tilde{X}_{i1}(\delta_1(\tilde{X}_{i1}; \beta_1)), \delta_1(\tilde{X}_{i1}; \beta_1), \tilde{X}_{i1}; \beta_2) \right) \right)$$

using any preferred maximization routine. For instance, Mebane and Sekhon (2009) describe a prepackaged global optimization solution that performs well for a wide range of difficult optimization problems. Under weak regularity conditions required for the Law of Large Numbers and for suitably large R, this two-step procedure yields the Bayesian optimal treatment regime within the class of rules $\mathcal{D}$. By estimating the optimal rule from posterior predictive simulations rather than from completed sample data or unbiased estimates of the parameters, we account for estimation uncertainty. Given an estimated optimal treatment regime, posterior inferences, such as the probability that the optimal treatment regime performs better than the status quo, follow the same logic as any other estimand.

Estimation of feasible sets follows that same exact steps, but iterated multiple times for different utility functions. To compute the set of feasible outcome means versus outcome variances, for instance, we can vary the weights λ_1 and λ_2 on a mean-variance utility function and optimize. Each set of weights returns a particular optimal rule. When applied to the posterior predictive outcomes, each rule returns a point on the feasible outcome set boundary. Alternatively, we can apply each rule to the completed sample data and compare these outcomes to the status quo regime.

1.6 Application: Educational Tracking

1.6.1 Data

I demonstrate these proposed methods through an application to ninth and tenth grade mathematics tracking in North Carolina public schools. Consistent with the two-period, binary treatment setup introduced earlier, I study a small component of the overall tracking problem: the decision to enroll in standard versus honors mathematics in ninth and tenth grade. I limit the sample to students enrolling in Algebra I in ninth grade. Roughly fifty percent of students take Algebra I in ninth grade, twenty percent enroll before ninth grade, and the remainder enroll after ninth grade. Students enrolling in earlier grades score significantly higher on the end-of-course Algebra I test despite being younger. The sample students are therefore neither the best nor the worst performers, and the tracking choice they face is the most common.

North Carolina's student testing program is one of the most comprehensive in the United States. Between third and eighth grade, students take annual end-of-year exams in reading and mathematics. After taking Algebra I, students take a standardized end-of-course exam and report their anticipated course grade. In tenth grade, students take further end-of-course exams, and, in some years, a comprehensive tenth grade mathematics assessment. I use longitudinal data, maintained by North Carolina Education Research Data Center (NCERDC), on students enrolled in tenth grade in the 2001-2002 school year.

Children's baseline covariates include gender, race, free or reduced lunch eligibility, score gain between seventh and eight grade, and eighth grade score. Intermediate outcomes include children's Algebra I test score and dichotomized anticipated grade. The final outcome is students' tenth grade comprehensive mathematics test score, which is available in 2001-2002. I normalize all variables to have mean zero and standard deviation one in each grade and drop students with incomplete records. Students are grouped into the honors track if they take either honors, advanced placement, or college placement mathematics. Because some schools are too small to offer comprehensive tracking options, I drop any school without students in each

track. The combined effect of dropping some schools and incomplete records leaves 24,112 students in 325 schools, compared to 44,696 ninth graders tested in Algebra I. While this is a significant reduction, most causes are plausibly missing at random, given the rich covariate information. The science, $\mathbf{S}_i$, includes 13 dimensions: four final outcomes (tenth grade scores), four intermediate outcomes (ninth grade scores and grades), and five baseline covariates (gender, race, free/reduced lunch eligibility, seventh to eighth gain score, and eighth grade score).

Consistent with intuition, students scoring higher in mathematics are significantly more likely to enroll in the ninth grade honors track, and there is clear evidence of dynamic selection into tenth grade honors (see appendex). Both the intermediate Algebra I end-of-course score and anticipated course grade have large and significant effects on the probability of enrolling in tenth grade honors mathematics. Given ninth grade track assignment, students performing better at the end of Algebra I are less likely to be demoted to a lower track and more likely to be promoted to a higher track. However, even given significant sorting, students enroll in standard and honors at all points in the test-score distribution in both ninth and tenth grade. There is significant overlap across tracks in the baseline covariate and intermediate outcome distributions.

1.6.2 Priors and computation

Given the large sample size, the exact prior specification is relatively unimportant. I use weak but proper priors primarily to ensure numerical stability of the unidentified scale parameters on the latent traits backing binary covariates such as gender. Specifically, with thirteen dimensions of the latent $\mathbf{S}_i^*$, I assume $\boldsymbol{a}_m = \mathbf{0}$, $\mathbf{B}_m = \mathbf{I}$,$a_V = 16$, $\mathbf{B}_V = \mathbf{I}$, $a_\nu = 2$, $b_\nu = 3$, $a_T = 16$, and $\mathbf{B}_T = \mathbf{I}$.

The random-effects multivariate normal model (1.14) is simple but quite flexible. The hierarchical Bayesian specification makes this flexibility feasible given limited data. With 325 schools and 13 dimensions to $\mathbf{S}_i^*$, there are $325 \cdot (13 + 13 \cdot (13-1)/2) = 29,575$ coefficients in $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$. In the application to educational tracking, the posterior mean for ν is roughly 80, which can be interpreted as 80 pseudo-observations

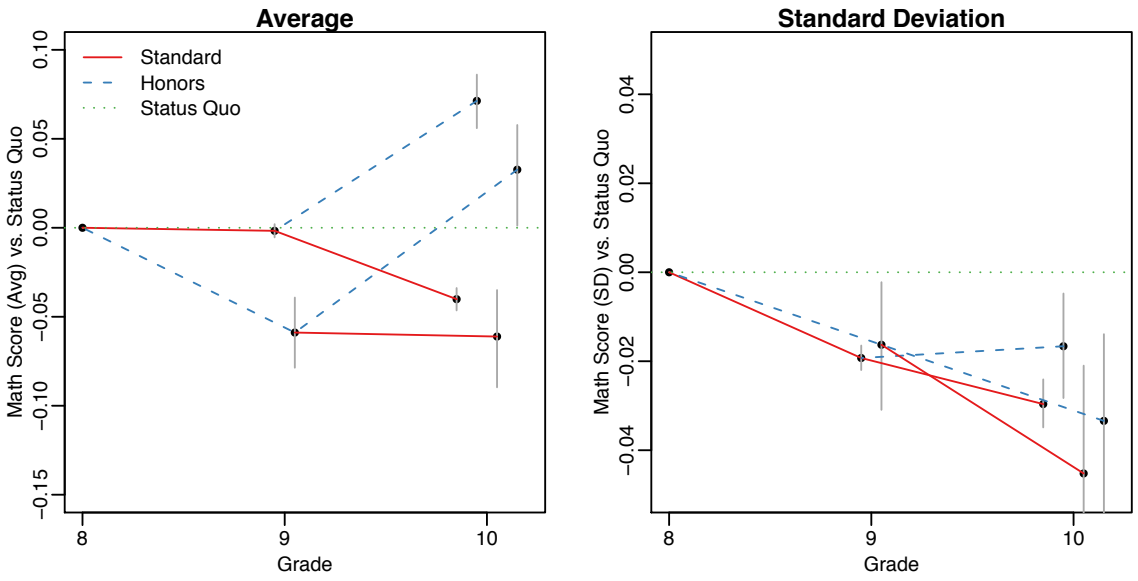


Figure 1.2: Math outcome for different tracking sequences. Dots give the posterior median of the average score (left) or standard deviation (right) under alternative tracking sequences. Results are compared to the observed status quo (horizontal line at zero). Bars give 90% credible intervals.

being added to each school when estimating $\mathbf{\Sigma}_j$. Posterior predictive checks also suggest the model fits the data well (see online supplement).

Computation follows the algorithm described in the appendix. The results use a MCMC chain of 10,000 iterations, retaining the last 9,000. Running the chain multiple times leads to similar results and the chain converges quickly. Due to the large portion of missing data and data augmentation steps, Gibbs sampling leads to significant autocorrelation in some parameters, although the main treatment effect parameters mix well. To compute the optimal treatment effect, I sample $R = 20,000$ future units, keeping the school assignment probabilities fixed at their sample means.

1.6.3 Basic results

Educators often discuss tracking policies in terms of both their impact on average achievement and inequality. I report both average effects and, in some cases, effects on the achievement standard deviation.

Figure 1.2 gives the sample-average treatment improvement (1.6) and posterior credible intervals for each tracking sequence compared to the status quo. A second panel gives the effect on the standard deviation. As can be seen, tracking students first into standard mathematics and then into honors leads to the highest average outcome. While standard tracking performs better in the short run than honors, by tenth grade standard/standard tracking outperforms only demotion from honors to standard. The honors/honors track experiences the largest one-year gain but still perform worse than standard/honors over both years ($p = 0.99$). Examining the entire tracking sequence therefore conveys a different picture than looking at each grade separately.

In terms of inequality, as measured by the outcome standard deviation, single-streaming students reduces inequality compared to the status quo, across all tracking sequences. The concern that status-quo tracking policies increase inequality appears to have some merit. Enrolling all students in standard/honors or honors/honors increases average performance while reducing inequality.

A major argument for tracking is that different courses are suitable for different students. Conditional results, reported in the appendix, are largely similar to Figure 1.2. One exception is that white and Asian students appear to suffer more than non-white/Asian students from demotion. Consistent with intuition, promotion to honors (standard/honors) is slightly more beneficial for top performers, whereas demotion to standard (honors/standard) is worse for top performers. Nevertheless, students of virtually all abilities (and enroll in Algebra I in ninth grade) perform best in the standard/honors track. Thus while significant heterogeneity exists, most of it is not relevant for optimal treatment choice.

1.6.4 Principal stratification on intermediate outcomes

Estimating causal effects stratified by intermediate outcomes is more subtle because observed intermediate outcomes are functions of earlier treatments. One approach to resolve this difficulty is to stratify instead on intermediate *potential* outcomes, which by definition do not depend on the treatment actually received. Doing so is

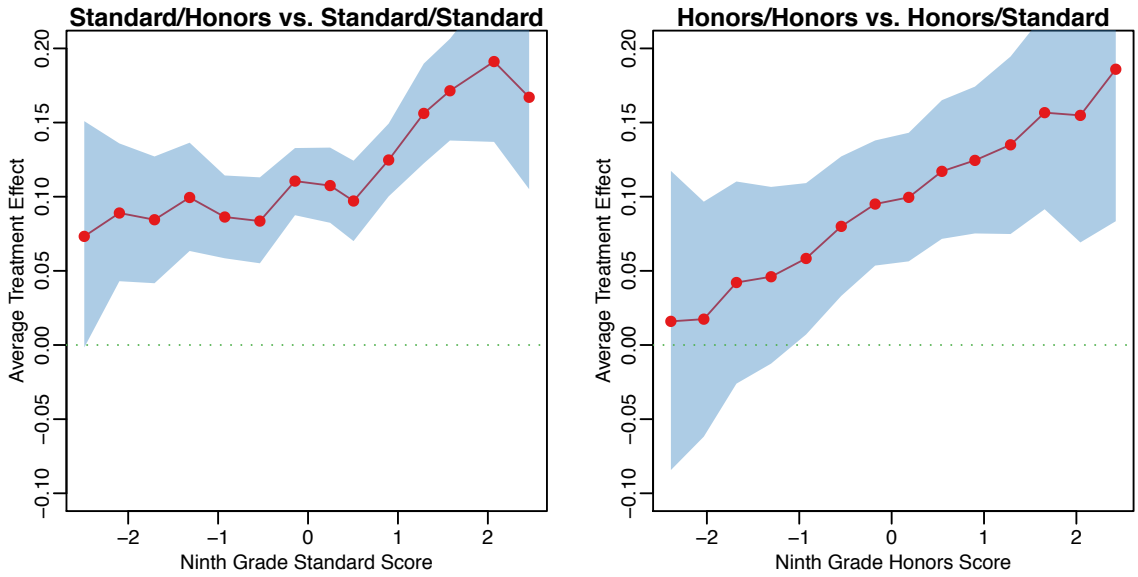


Figure 1.3: Conditional sample-average treatment effect by intermediate potential performance. Dots give the posterior binned (conditional) sample-average treatment effects by intermediate potential ninth grade performance for two treatment sequences contrasts. Shaded areas give pointwise 90% credible intervals. The x-axis is potential ninth grade performance, including all units within the sample, not observed performance.

an example of principal stratification (Frangakis and Rubin, 2002), applied to continuous intermediate outcomes. Estimating causal effects stratified by intermediate potential outcomes is often of direct interest to policymakers; it allows them to impose counterfactual policies in earlier periods but still explore heterogeneity for later treatments.

Figure 1.3 shows the conditional sample-average treatment effect of promotion (standard/honors vs. standard/standard) and demotion (honors/standard vs. honors/honors) conditional on intermediate potential achievement. By conditioning on the entire sample's intermediate potential performance given standard (left) or honors (right) in ninth grade, Figure 1.3 differs from the alternative of treating ninth grade as the baseline and allowing for heterogeneity by the observed ninth grade track and achievement. In contrast to that approach, Figure 1.3 asks: had *all* students been assigned to standard (left) or honors (right) in ninth grade, how would

Analysis Assumptions	Students' Mathematics Track in Ninth and Tenth Grade			
	Standard, Standard	Standard, Honors	Honors, Standard	Honors, Honors
Intermediate selection and effects	-0.04	0.07	-0.06	0.03
	[-0.05 , -0.04]	[0.06 , 0.08]	[-0.09 , -0.04]	[0 , 0.05]
No intermediate selection	-0.05	0.13	-0.08	0.08
	[-0.06 , -0.05]	[0.11 , 0.14]	[-0.11 , -0.06]	[0.05 , 0.11]
No intermediate effects	-0.04	0.11	-0.11	0.21
	[-0.05 , -0.03]	[0.06 , 0.12]	[-0.12 , -0.09]	[0.14 , 0.27]

Table 1.1: Estimated outcomes ignoring intermediate selection or effects. Cells contain posterior modes for the average effect and brackets contain 90% credible intervals, compared to the status quo. The first row accounts for dynamic selection and intermediate effects. The second row drops intermediate ninth grade math scores and grades, and therefore ignores dynamic selection. The third row assumes that there are no intermediate effects on math scores or grades, and therefore includes these variables as baseline controls.

honors versus standard tracking in tenth grade affect students' final outcomes given their intermediate, ninth grade, performance? As can be seen, regardless of the hypothetical ninth grade tracking assignment, higher-achieving students at the end of ninth grade gain more from being assigned into honors mathematics in tenth grade than lower-achieving students, although all benefit.

1.6.5 Ignoring dynamic selection or intermediate effects

The preceding results allow for both dynamic selection and effects on intermediate outcomes. Table 1.1 shows the results from an analysis that assumes either no dynamic selection or no intermediate effects. Both assumptions are unreasonable. Students are promoted and demoted after ninth grade on the basis of their intermediate performance, even controlling for their baseline achievement, and intermediate performance is not predetermined—it directly depends on students' ninth grade tracking assignment.

The results demonstrate the problem with analyzing time-varying treatments using standard tools. The first line shows the estimated effects accounting for dynamic

selection and allowing for intermediate effects, as in Figure 1.2. The second row assumes no dynamic selection by dropping intermediate outcomes. Ignoring dynamic selection doubles the estimated treatment effect of standard/honors and honors/honors compared to standard/standard. This makes sense since high performing students are filtered from standard to honors over time. Dynamic selection therefore biases the estimated effect of honors upward. The third row assumes no intermediate effects by treating post-treatment intermediate outcomes as baseline covariates. Doing so also induces significant biases. For instance, the effect of honors/honors compared to the status quo jumps from 0.03 to 0.21. Figure 1.2 provides intuition for this result: for the sample considered, honors in ninth grade has a negative effect compared to standard, which we falsely mitigate by assuming $X_{i2}(0) = X_{i2}(1) = X_{i2}$.

1.6.6 Optimal treatment and feasible outcome sets

While the precise objectives of educators are complex, a major goal, particularly given the rise of test-based accountability systems, is to increase average performance. With this goal in mind, I explore the design of optimal dynamic treatment regimes that maximize average tenth grade achievement.

I consider three types of treatment regimes. The simplest is a dynamic cutoff regime. Cutoff rules are easy to understand and implement. The dynamic cutoff rule assigns students to honors if their last mathematics test score exceeds a cut-point defined separately for each grade and track. I also explore static and dynamic linear index regimes that assign students to tracks if a linear index of the observed data, including a constant term, exceeds zero. In contrast to the static regime, the dynamic regime includes intermediate outcomes in the tenth grade decision function. Following the approach described in Section 1.5.2, I select each rule by simulating 20,000 hypothetical future units and estimating the optimal treatment regime parameters using an off-the-shelf global maximization routine (Mebane and Sekhon, 2009).

Tables 1.2 and 1.3 summarize characteristics of the resulting track assignments if we apply the estimated optimal treatment regimes retrospectively to the sample data. In Table 1.2 each cell contains the estimated fraction of students assigned to each

	Standard, Standard	Standard, Honors	Honors, Standard	Honors, Honors
Status quo regime	0.64	0.20	0.06	0.10
Dynamic cutoff regime	0.00	1.00	0.00	0.00
Static linear index regime	0.00	0.86	0.01	0.13
Dynamic linear index regime	0.01	0.87	0.04	0.08

Table 1.2: Fraction of student assigned to each track by treatment regime. Cells give the posterior median of the fraction of (sample) students enrolled in each track under different treatment regimes. Posterior standard errors for dynamic rules that depend on potentially unobserved intermediate outcomes are less than 0.01 and therefore suppressed.

	Standard, Standard	Standard, Honors	Honors, Standard	Honors, Honors
Fraction Female	0.00	0.56	0.04	0.16
Fraction White/Asian	0.00	0.63	0.54	0.81
Fraction Free/Reduced Lunch	0.49	0.16	0.03	0.01
Average Seventh to Eighth Score Gain	0.00	0.17	-1.21	-1.23
Average Eighth Grade Score	-1.60	0.09	-1.12	-0.27
Average Ninth Score Grade Score (Standard)	-1.36	0.03	-0.72	0.10
Average Ninth Score Grade Score (Honors)	-1.52	-0.02	-0.71	0.03
Fraction Ninth Grade A or B (Standard)	0.26	0.50	0.02	0.55
Fraction Ninth Grade A or B (Honors)	0.00	0.47	0.42	0.44

Table 1.3: Characteristic by track under dynamic linear index regime. Cells give posterior means of the sample characteristics for students in each track when assigned using the estimated optimal dynamic linear index regime. Standard errors, which are extremely small, are suppressed.

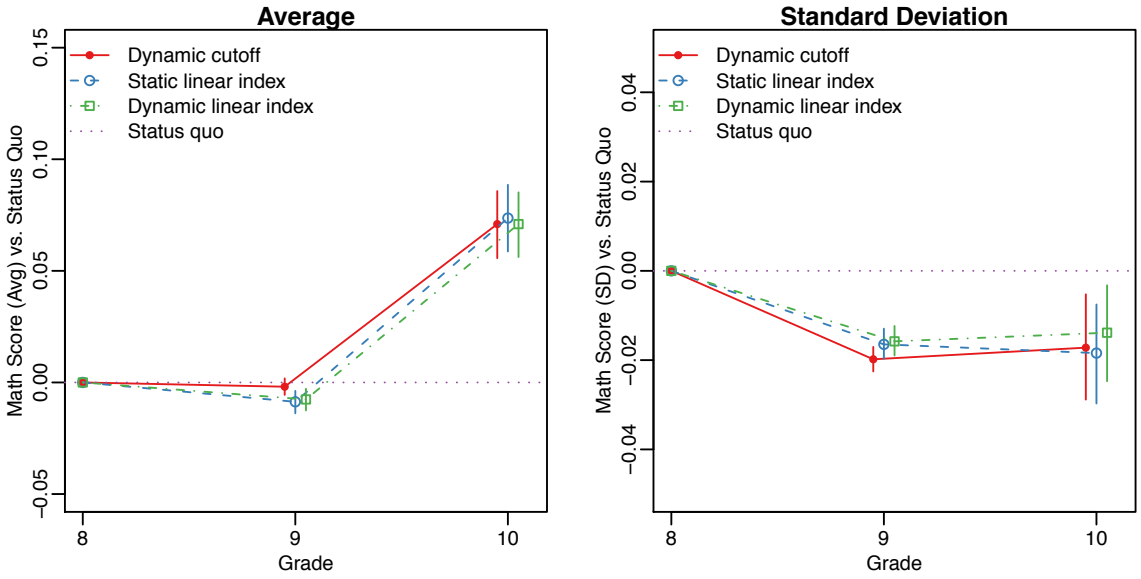


Figure 1.4: Outcomes under alternative optimal treatment regimes. Dots give the posterior median of the mean score (left) or standard deviation (right), compared to the status quo, under alternative optimal treatment regimes. Bars give 90% credible intervals. Effects are sample treatment regime effects.

track. The dynamic cutoff rule assigns virtually all students to the standard/honors track. This result is consistent with the standard/honors tracking sequence consistently outperforming other tracking paths. By comparison, the status quo has significantly less track mobility—only 30 percent of students switch tracks between ninth and tenth grade. Under both the static and dynamic linear index rule, the standard/honors track also receives over 85 percent of students. Table 1.3 shows the characteristics of students assigned to each track under the optimal dynamic linear index regime. Students assigned to honors/standard and honors/honors have lower baseline scores than the dominant standard/honors track.

Figure 1.4 shows the main result: the effect of each optimal tracking regime on average achievement and the standard deviation of achievement. All three regimes outperform the status quo tracking regime by roughly 0.07 standard deviations, while slightly lowering the standard deviation of achievement. Within the sample, the static linear regime performs best, although the dynamic regime should always obtain a

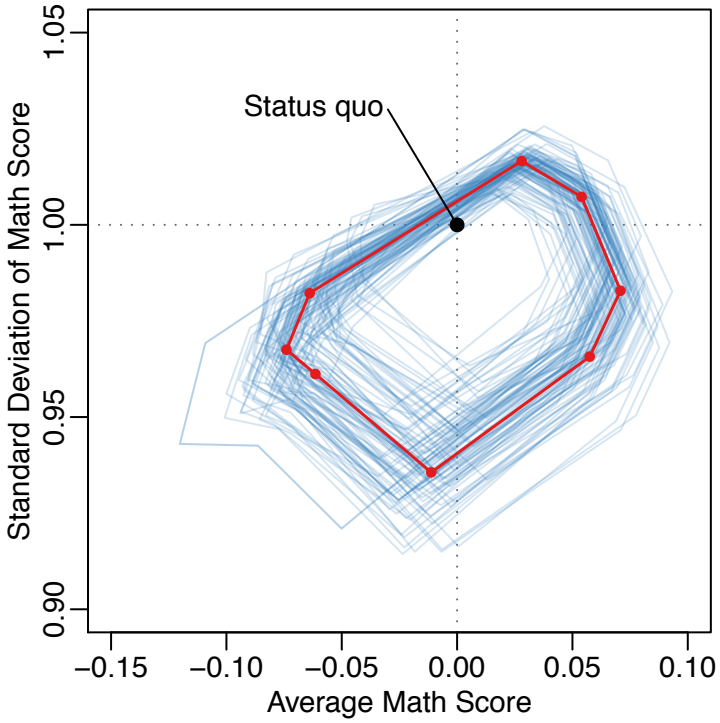


Figure 1.5: Feasible math outcomes under a dynamic cutoff regime. The dark line outlines the posterior mean for the (sample) feasible outcome set given a dynamic cutoff regime. Grey lines give posterior draws of feasible outcome sets. For computational reasons, sets are computed at eight points. The status quo point gives the observed outcome under the status quo track assignment policy. In contrast to population feasible sets, sample feasible sets are not necessarily convex, as can be seen in some posterior draws.

better population outcome than the static regime. Virtually all the gains come from tracking students to the standard/honors sequence.

Finally, Figure 1.5 shows the feasible outcome set under a dynamic cutoff regime. I compute the set by choosing eight different weights on the mean and variance preferences and maximizing utility. The boundaries therefore represent an eight-point approximation of the feasible set. To facilitate a comparison to the status quo, I apply the population-optimal regimes to the sample data, which can lead to non-convex feasible sets for some posterior draws. Each light line represents a posterior draw of the sample tenth grade achievement mean and standard deviation under the

eight regimes. The dark lines and dots give the posterior expectation. If policymakers favor lower variance and higher mean, then the lower-right part of the plot represents potential improvements over the status quo. As can be seen, there is significant room for improvement both in terms of the average and standard deviation of performance. The status quo outcome falls in the undesirable upper-left portion of the feasible outcome set. There is no binding equity-efficiency trade off and current tracking policies exacerbate inequality.

1.7 Conclusion

Time-varying treatments and dynamic treatment regimes lie at the heart of many different disciplines. Yet standard program evaluation methods are not designed to handle dynamic selection into treatment and intermediate causal effects. This paper extends the framework for missing data and causal inference introduced by Rubin (1976, 1978) and Bayesian decision theory to study dynamic treatment regimes and treatment choice. I propose a model on the joint distribution of potential outcomes and baseline covariates that can handle continuous and discrete variables and grouped data. While the conceptual framework is general, there is considerable room for research on models suitable for applications with many intermediate variables and/or many time points, and methods to compare different model specifications and to assess the sensitivity of results to model and prior choices.

Applied to educational tracking, the proposed methods provide inference for central quantities in the policy debate: outcomes under different tracking regimes, the optimal tracking regime, student mobility between tracks, and the tradeoff between equity and efficiency in student outcomes. The results suggest room for improving tracking policies, including rules which simultaneously increase achievement while reducing inequality. However there are caveats to these results. They apply only to the sample of students that enroll in Algebra I in ninth grade. The analysis also assumes that peer effects associated with rearranging students are negligible and that teachers do not adapt when faced with different students. Richer models of classroom

behavior, which extend beyond the time-varying dimension of education, may further improve the design of optimal tracking rules.

The Bayesian approach to dynamic treatment regimes has a number of appealing characteristics. A key idea is that sequential unconfoundedness is simply a particular ignorable treatment mechanism. After carefully defining the missing and observed data, Bayesian inference follows the same steps as multiple imputation procedures developed outside of the dynamic treatment context. Bayesian hierarchical models also offer the potential to partially pool information across treatment paths or groups. Pooling information is often critical given many treatment paths or in applications such as education where grouped structures are common. Finally, the proposed methods integrate over all sources of uncertainty when performing treatment choice and can address more complex questions such as estimation and inference for feasible outcome sets. Extending causal analysis to dynamic treatment regimes has the potential to deepen understanding of causal mechanisms and improve policy and practice across different disciplines.

1.A Appendix

1.A.1 Inference given randomized treatment or no intermediate effects

All ignorable treatment rules lead to the same posterior inference. This result implies that a full-information Bayesian analysis proceeds equivalently regardless of whether treatment is sequentially unconfounded or completely randomized. Nevertheless, completely randomized treatments do offer the possibility of limited-information analysis that ignores intermediate outcomes. For example, given randomization conditional on baseline characteristics, $f(\mathbf{W}_2, \mathbf{W}_1 \mid \mathbf{S}) = f(\mathbf{W}_2, \mathbf{W}_1 \mid \mathbf{X}_1)$, it becomes possible to ignore intermediate outcomes. While the full Bayesian paradigm demands conditioning on all available information, ignoring intermediate outcomes can simplify the analysis. The problem reduces to a point-treatment setting with four treatments.

Specifically,

$$f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}, \mathbf{W}_2, \mathbf{W}_1, \mathbf{X}_1) \propto f(\mathbf{Y}(0,0), \mathbf{Y}(0,1), \mathbf{Y}(1,0), \mathbf{Y}(1,1), \mathbf{X}_1). \tag{1.15}$$

In applications where randomization can be justified, time-varying treatments can therefore be estimated with the standard tools used to evaluate point-treatment causal effects. However, given dynamic selection based on intermediate outcomes, which almost always occurs in sequential observational settings, such methods will fail; one can no longer ignore the intermediate outcomes and still obtain (1.15).

Another special case arises when intermediate outcomes exist, but the science of the problem suggests that treatment has no causal effect on intermediate outcomes. That is, when we know a priori that $\mathbf{X}_2(0) = \mathbf{X}_2(1)$. For instance, it may be plausible to assume that tracking policies have no effect on time-varying parental employment, but still potentially allow for parental employment to influence tracking decisions and be correlated with potential outcomes. Full-information analysis proceeds by treating intermediate outcomes as additional baseline covariates. Under ignorable treatment, and assuming no intermediate causal effects,

$$f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}, \mathbf{W}_2, \mathbf{X}_2, \mathbf{W}_1, \mathbf{X}_1) \propto f(\mathbf{Y}(0,0), \mathbf{Y}(0,1), \mathbf{Y}(1,0), \mathbf{Y}(1,1), \mathbf{X}_2, \mathbf{X}_1).$$

This is equivalent to point-treatment analysis with four treatments and $\mathbf{X}_2$ and $\mathbf{X}_1$ as baseline covariates. If, as is typically the case, treatment does affect intermediate outcomes, then treating intermediate outcomes as additional baseline covariates will lead to incorrect inference. Therefore, except in limited special cases, we cannot analyze time-varying treatments using the same tools developed for point-treatment settings.

1.A.2 Data and selection

Table 1.4 gives summary statistics for students by each tracking sequence. Table 1.5 tests dynamic selection into tracks using a logit model of track assignment in ninth

Measure	Students' Mathematics Track in Ninth and Tenth Grade			
	Standard, Standard	Standard, Honors	Honors, Standard	Honors, Honors
Female	0.564	0.55	0.549	0.559
White or Asian	0.693	0.766	0.674	0.741
Free or Reduced Lunch	0.169	0.124	0.167	0.155
Grade 7 to 8 Math Gain Score	-0.016	0.086	-0.069	-0.025
	[1.013]	[0.965]	[0.993]	[0.977]
Grade 8 Math Score	-0.143	0.377	-0.126	0.241
	[0.947]	[1.035]	[0.941]	[1.056]
Grade 9 Math Score	-0.132	0.386	-0.215	0.214
	[0.941]	[1.037]	[0.98]	[1.06]
Grade 9 Algebra I Grade (A or B)	0.486	0.724	0.491	0.652
Grade 10 Math Score	-0.151	0.414	-0.179	0.257
	[0.942]	[1.026]	[0.955]	[1.045]
N	15362	4744	1529	2477

Table 1.4: Sample summary statistics by mathematics track. Cells contain means and brackets contain standard deviations. All continuous variables have been normalized to have mean zero and standard deviation one for each grade within the sample.

and tenth grade on past observables. These models describe the status quo treatment assignment mechanism.

1.A.3 Model checking

While ignorable treatment assignment is untestable without additional assumptions, it is useful to check the sensitivity of inferences to model modifications. Zhang et al. (2009), for instance, consider how Box-Cox transformations of wage rates influences the log-likelihood and average treatment effect of a training program. In multivariate applications, mixture models, copula-based multivariate distributions, conditional factorizations, or different random effects structures provide other avenues to explore.

One can often motivate model modifications by comparing posterior predictive replications under a starting model to the observed data (Rubin, 1984; Gelman et al., 1996). That is, does the model generate data that looks similar to the observed data?

	Ninth Grade Honors	Tenth Grade Honors given Ninth Grade Standard	Tenth Grade Honors given Ninth Grade Honors
Female	0.01	-0.01	0.07
	(0.04)	(0.04)	(0.07)
White or Asian	-0.09	-0.08	0.08
	(0.04)	(0.04)	(0.08)
Free or Reduced Lunch	0.05	-0.16	0.18
	(0.05)	(0.05)	(0.1)
Grade 7 to 8 Math Gain Score	-0.11	-0.03	-0.01
	(0.02)	(0.02)	(0.04)
Grade 8 Math Score	0.17	0.31	0.12
	(0.02)	(0.03)	(0.05)
Grade 9 Math Score		0.2	0.25
		(0.03)	(0.05)
Grade 9 Algebra I Grade (A or B)		0.65	0.36
		(0.04)	(0.07)
Intercept	-1.57	-1.54	0.15
	(0.04)	(0.05)	(0.09)
N	24112	20106	4006

Table 1.5: Logit treatment selection model. Coefficients are expressed in logits from a model with weakly informative Cauchy prior (Gelman et al., 2008). Parentheses give the posterior standard deviations.

Such posterior predictive checks can be formalized to return posterior predictive p-values for discrepancy statistics or used graphically to motivate model modification. With missing data, standard posterior predictive checks require modeling the missing data mechanism. As an alternative, Gelman et al. (2005) propose comparing completed data, which includes both imputed and observed data, with replicated data. Completing the data first avoids the need to model the missing data mechanism. Moreover, the completed data $\mathbf{S}$ is generally easier to interpret in terms of the original model than the observed $\mathbf{S}^{\text{obs}}$. However, working with completed rather than observed data may reduce power since both the replicated and completed data are based on the model.

Following this strategy, Figure 1.6 show graphical posterior predictive checks of the marginal distribution of final potential outcomes, which are the most important model components. Specifically, Figure 1.6 uses twenty completions and replications from $f(\mathbf{S}, \mathbf{S}^{\text{rep}} \mid \mathbf{Z})$, where $\mathbf{S}$ denotes a $N \times 13$ completed data matrix for the observed sample and $\mathbf{S}^{\text{rep}}$ denotes an $N \times 13$ posterior predictive replicated dataset, holding the school assignment probabilities fixed. If the model is correct, the quantile-quantile (QQ) plots of completed and replicated achievement measures should fall along the diagonal. As the plots show, the model fits the observed data, with no large deviations from the expected distribution. The one exception is a small but statistically significant deviation in the lower part of the distribution for the $\mathbf{Y}(0,0)$ outcome. Similar posterior predictive checks for the conditional means and variances show no systematic deviations, except for artifacts caused by ignoring test-score rounding.

1.A.4 Extended results

Table 1.6 presents the parameters of the estimated optimal treatment regimes. The dynamic cutoff rule assigns only the very best students—those with eighth grade scores 3.3 standard deviations above the average—to honors in ninth grade. Given assignment to standard in ninth grade, the optimal cutoff rule assigns any student scoring above -6.6 to honors in tenth grade. Interpreting the linear and dynamic

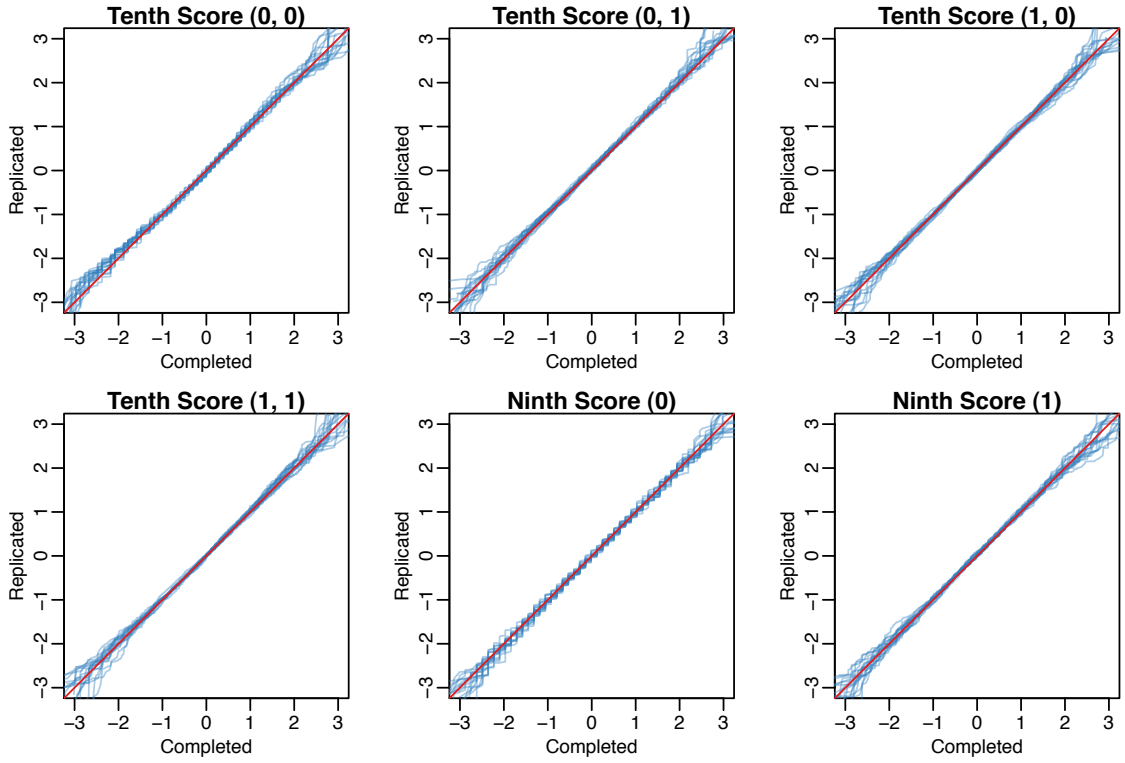


Figure 1.6: Posterior predictive checks of the marginal outcome distributions. QQ-plots of 20 completed intermediate and final potential outcomes for different treatment sequences versus posterior predictive replications. Under the model the QQ-lines should fall along the diagonal.

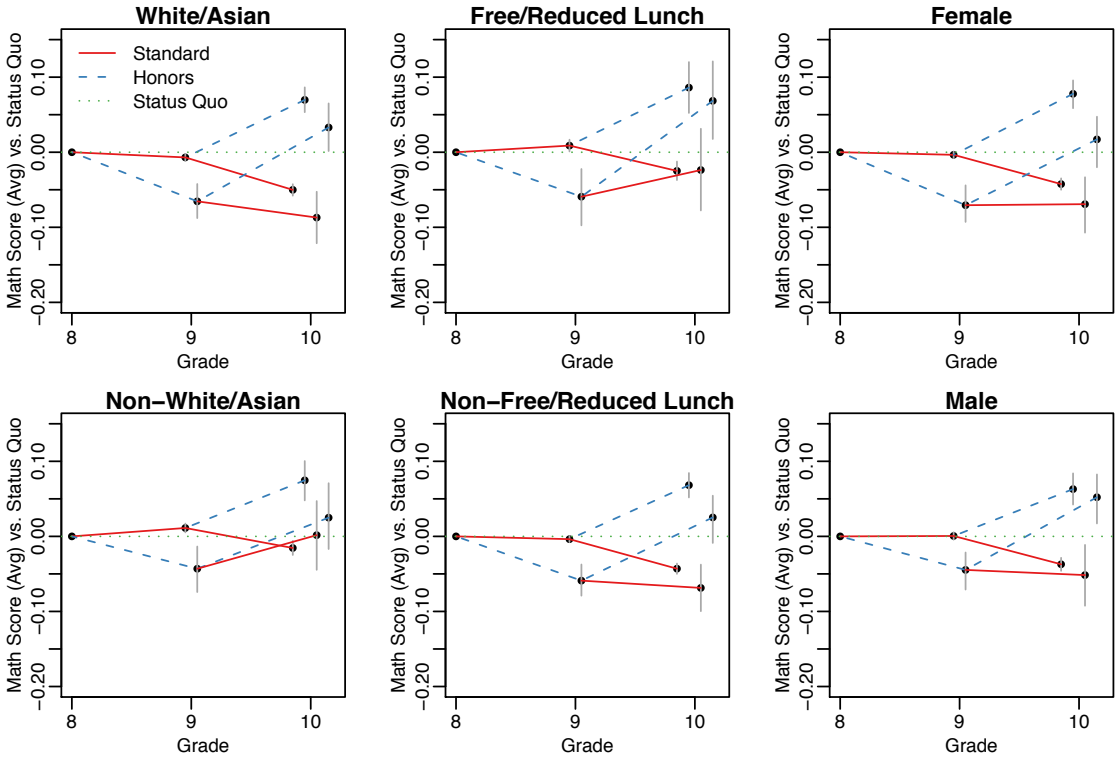


Figure 1.7: Math outcomes by subpopulation for different tracking sequences.

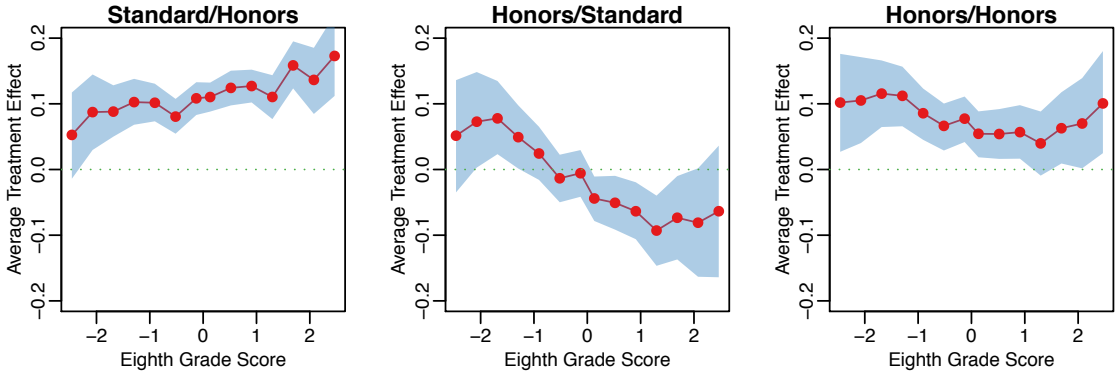


Figure 1.8: Conditional sample-average treatment effect vs. standard/standard by baseline performance. Dots give the posterior binned sample means of tenth grade scores under different tracking sequences compared to enrolling in standard in both periods (horizontal line at zero). Shaded areas give pointwise 90% credible intervals.

	Dynamic Cutoff Regime			Static Linear Index Regime		Dynamic Linear Index Regime	
	Grade 9	Grade 10	Grade 10	Grade 9	Grade 10	Grade 9	Grade 10
Constant (Intercept)				-2.9	9.6	-3.9	2.7
Female				-4.6	-3.2	-6.1	5.4
White or Asian				-4.9	5.4	1.1	3.7
Free or Reduced Lunch				-2.7	3.3	-6.8	0.3
Grade 8 Math Score	3.3			-8.8	3.6	-4.7	0
Grade 7 to 8 Math Gain Score				1.6	-0.4	-1.1	1.9
Grade 9 Math Score		-6.6	2.1				0.3
Grade 9 Algebra I Grade (A or B)							6.5
Honors in Ninth Grade		No	Yes		-1.4		-5.2

Table 1.6: Alternative optimal treatment regimes. Cells give the coefficients or cutoffs of the estimated optimal treatment regimes. The dynamic cutoff regime assigns students to honors if their most recent mathematics score exceeds a particular cutoff defined for each track. The static and dynamic linear index regimes constructs a linear index based on each students measured outcomes and previous track, and assigns units to honors if the index exceeds zero. Indices are created separately for each grade, but not each track.

index rules is more difficult due to the correlation between covariates. The final column shows clear positive sorting on the intermediate performance.

1.A.5 Complications: peer effects and non-ignorable treatment

In some applications, such as educational tracking, the assumption of no interaction between units implicit in the potential outcome notation may be suspect. For instance, researchers often assume that students' classroom peers have a causal impact on performance. Conceptually, incorporating peer effects simply requires an expansion of potential outcome space. With interaction between units, each unit has potential outcomes associated with every possible combination of treatment assignments for the entire group. For instance, if $\overline{\mathbf{w}} = ((w_{11}, w_{12}), \ldots, (w_{N1}, w_{N2}))$ denotes a potential treatment allocation for all units, then units have potential outcomes $Y_i(\overline{\mathbf{w}})$, which incorporate the potential influence of peers. Given two periods and a binary treatment, this leads to 4^N potential outcomes for each unit.

To make estimation practical, we clearly must limit the potential outcome space in some way. In the context of education, students may be affected only by peers in the same classroom and through particular functions of their peer's characteristics. One approach might specify a measure of peer quality p and assume that peer effects are linearly additive, i.e., $Y_i(w_1, w_2, p) = Y_i(w_1, w_2) + \alpha p$. Incorporating peer effects nevertheless can be quite complex in practice and is left as a potential future extension.

A second concern, but one that is inherently untestable, is that treatment assignment is confounded. While the rich longitudinal information available in administrative data mitigates this concern, policymakers may rely on unobserved information related to the missing potential outcomes to determine treatment. Absent ignorable treatment, Bayesian analysis requires modeling the treatment assignment mechanism. In practice, many of the ideas used in point-treatment settings may be usefully extended to sequential treatments. For instance, in applications where a credible exogenous source of variation exists, instrumental variable approaches similar to Imbens

and Rubin (1997), Hirano et al. (2000), and Frangakis and Rubin (2002) may be applicable. However, further work is needed to extend these strategies to sequential settings.

1.A.6 Posterior inference using MCMC

We can explore the posterior distribution of the parameters and missing data using Markov Chain Monte Carlo (MCMC) methods. The sampling procedure proceeds by iterating between data augmentation steps for the missing potential outcomes and latent science, and posterior draws from the parameters given the completed data, along with updates of the priors. If desired, posterior predictive simulations can be made conditional on each parameter draw. The following describes a simple MCMC algorithm for posterior inference in the random-effects multivariate normal model.

Step 0: Initialize the missing data and parameters.

Initialize the missing potential outcomes, latent variables, and parameters at some reasonable value; for instance, by drawing values from the prior.

Step 1: Augment the missing potential outcomes.

Let superscripts $\mathbf{S}_i^1$ denote the unobserved potential outcomes and $\mathbf{S}_i^2$ denote the observed data of the latent $\mathbf{S}_i^*$ for unit i. Then a Gibbs data augmentation step for the missing potential outcomes follows the properties of the multivariate normal

$$f(\mathbf{S}_i^1 \mid \cdots) = \mathcal{N}\left(\boldsymbol{\mu}_j^1 + \boldsymbol{\Sigma}_j^{12}\left(\boldsymbol{\Sigma}_j^{22}\right)^{-1}(\mathbf{S}_i^2 - \boldsymbol{\mu}_j^2), \boldsymbol{\Sigma}_j^{11} - \boldsymbol{\Sigma}_j^{12}\left(\boldsymbol{\Sigma}_j^{22}\right)^{-1}\boldsymbol{\Sigma}_j^{21}\right),$$

where superscripts define the partitioned matrices associated with the observed and missing components, and subscripts indicate the school assignment j for child i. In this notation, the meaning of superscripts varies by unit, depending on the pattern of missing data.

Step 2: Augment the discrete components with their latent representations.

Draw $\mathbf{S}_i^{d*}$ from a truncated multivariate normal with mean and variance following the properties of the partitioned conditional multivariate normal, similar to Step 1, and truncation in each dimension k as $(\infty, 0]$ if $\mathbf{S}_{ik}^d = 0$ and $(0, \infty)$ if $\mathbf{S}_{ik}^d = 1$. A simple sampling method, although not the most efficient, takes multiple univariate Gibbs steps, each from a truncated normal drawn by the inverse-CDF method.

A difficulty in latent variable formulations for discrete data are that the variances are not identified. That is, the observed data are invariant to scale transformation of the latent variables and parameters. In theory, this situation poses no difficulty to Bayesian methods, since we can resolve identification issues by placing proper priors on all the parameters. In practice, some care must be taken to avoid numerically unstable estimates. Following Edwards and Allenby (2003), I use proper priors on all the parameters and post-process the data to restrict the variances to one for the discrete components.

Step 3: Update the parameters $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$.

Sample $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ from their full conditionals. Within school j,

$$f(\boldsymbol{\mu}_j \mid \cdots) = \mathcal{N}\left(\left(\mathbf{V}^{-1} + N_j\boldsymbol{\Sigma}_j^{-1}\right)^{-1}\left(\mathbf{V}^{-1}\boldsymbol{m} + \boldsymbol{\Sigma}_j^{-1}\sum_i \mathbf{S}_i^*\right), \left(\mathbf{V}^{-1} + N_j\boldsymbol{\Sigma}_j^{-1}\right)^{-1}\right),$$

$$f(\boldsymbol{\Sigma}_j \mid \cdots) = \mathcal{IW}\left(\nu + N_j, \mathbf{T} + \sum_i (\mathbf{S}_i^* - \boldsymbol{\mu}_j)(\mathbf{S}_i^* - \boldsymbol{\mu}_j)'\right),$$

where N_j is the number of children in school j, and the summations are taken for children within school j.

Step 4: Update the priors m, V, ν and T.

Sample priors $\mathbf{m}$, $\mathbf{V}$, and $\mathbf{T}$ from their full conditionals:

$$f(\mathbf{m} \mid \cdots) = \mathcal{N}\left(\left(\mathbf{B}_m^{-1} + J\mathbf{V}^{-1}\right)^{-1}\left(\mathbf{B}_m^{-1}\mathbf{a}_m + \mathbf{V}^{-1}\sum_{j=1}^{J}\boldsymbol{\mu}_j\right), \left(\mathbf{B}_m^{-1} + J\mathbf{V}^{-1}\right)^{-1}\right),$$

$$f(\mathbf{V} \mid \cdots) = \mathcal{IW}\left(a_V + J, \mathbf{B}_V + \sum_{j=1}^{J}(\boldsymbol{\mu}_j - \mathbf{m})(\boldsymbol{\mu}_j - \mathbf{m})'\right),$$

$$f(\mathbf{T} \mid \cdots) = \mathcal{W}\left(a_T + J\nu, \left(\mathbf{B}_T^{-1} + \sum_{j=1}^{J}\boldsymbol{\Sigma}_j^{-1}\right)^{-1}\right),$$

where J is the number of schools or groups. Finally, we can update ν by taking a Metropolis-Hastings or slice-sampling step using

$$f(\nu \mid \cdots) \propto \ln\mathcal{N}\left(\nu - K - 1 \mid a_\nu, b_\nu\right) \cdot \sum_{J=1}^{J}\mathcal{IW}\left(\boldsymbol{\Sigma}_j \mid \nu, \mathbf{T}\right),$$

where $\ln\mathcal{N}(\cdot \mid \cdot)$ and $\mathcal{IW}(\cdot \mid \cdot)$ represent the log-normal and inverse-Wishart density functions, and K denotes the dimension of $\boldsymbol{\Sigma}_j$.

Step 5: Draw posterior predictive simulations as needed.

If necessary, draw posterior predictive simulations for future units from the model. This step requires a multivariate normal draw of the latent science, $\mathbf{S}_i^* \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, coupled with a transformation for the discrete components. For simplicity, I assume the assignment probability to group j matches the sample data.

By standard MCMC results, iterating these steps will cause the chain to converge to a stationary distribution equal to the desired posterior distribution. After a suitable burn-in period, we can then estimate any of the causal estimands introduced in Section 2 using either the completed data from Step 1 or the posterior predictive simulations from Step 5. In a multivariate normal model, the parameters $\boldsymbol{\mu}_j$ associated with the potential outcomes give the posterior population mean under each treatment and can be used to estimate population-average effects directly. For other models, this straightforward mapping between parameters and causal effects may not always apply, in which case posterior predictive simulations for hypothetical future units provides an alternative.

Chapter 2

Regression Discontinuity Design with Multiple Forcing Variables[1]

Regression discontinuity designs identify causal effects by exploiting treatment assignment rules that are discontinuous functions of underlying covariates. In the standard regression discontinuity design setup, the probability of treatment changes discontinuously if a scalar covariate exceeds a cutoff. We consider more complex treatment assignment rules that generate a treatment boundary. Leading examples include education policies where treatment depends on multiple test scores and spatial treatment discontinuities arising from geographic borders. We give local linear estimators for both the conditional effect along the boundary and the average effect over the boundary, and a consistent estimate for the variance of the average effect based on the nonparametric delta method. For two-dimensional RD designs, we derive an optimal, data-dependent, bandwidth selection rule for the conditional effect. We demonstrate these methods using a summer school and grade retention example.

[1]An modified version of this chapter may appear with Guido Imbens. Thanks go to Brian Jacob and Lars Lefgren, and the Chicago Public Schools, for making their data available. Raymond Guiteras and Brigham Frandsen provided the actual final data we use and valuable clarifying documentation.

2.1 Introduction

Regression discontinuity (RD) designs identify causal effects by exploiting treatment assignment rules that are discontinuous functions of underlying covariates (Thistlethwaite and Campbell, 1960; Hahn et al., 2001). In the classic RD setup, the probability of treatment changes discontinuously if a scalar characteristic falls above or below a cutoff. But treatment assignment rules can be more complex and depend on a vector of covariates. In education settings, for instance, children often must pass multiple subject exams to avoid summer school, advance to the next grade, or graduate (e.g., Jacob and Lefgren, 2004; Martorell, 2005; Matsudaira, 2008; Papay et al., 2008). Likewise, public policies can differ sharply across political, administrative, or geographic borders defined by latitude and longitude (e.g., Holmes, 1998; Black, 1999; Pence, 2006; Lalive, 2008; Dell, 2010; Gerber et al., 2010).

Starting with Papay et al. (2009), a number of researchers have proposed methods to extend RD design to applications with multiple forcing variables (e.g., Wong et al., 2010; Reardon and Robinson, 2010). As highlighted by this work, RD designs with multiple forcing variables identify the conditional treatment effect at every point along the treatment boundary rather than at a single point. In many applications, researchers may wish to explore this heterogeneity.

Motivated by this initiating work on boundary RD designs, this paper studies a particular approach: local linear estimation of RD designs that generate a treatment boundary. Local linear estimation has proven successful in scalar contexts but a number of open issues remain given multiple forcing variables. To start, we give estimators for two parameters identified by boundary RD designs. The first is a local linear estimator for the conditional treatment effect at any point on the treatment boundary. This estimator is a generalization of the local linear estimator proposed by Hahn et al. (2001) for the scalar case. We consider both sharp designs, where treatment is solely determined by the forcing variables, and fuzzy designs, where the treatment probability changes discontinuously across the boundary.

Second, we also discuss estimation and inference for the average effect over the boundary. The common approach of using the distance to the nearest boundary in a

scalar RD setup gives a consistent estimate for an average effect along the boundary, although the weighting scheme may differ from explicit integration along the boundary for some designs. In most applications converting boundary RD designs to a scalar RD design works well. However, we also explore directly averaging the conditional effect over the boundary. At the expense of additional complexity, averaging explicitly can have better finite sample properties. We give an estimator for the variance of the explicitly integrated average effect based on the nonparametric delta method.

When interest lies only in the average effect along the boundary, we recommend converting boundary RD designs into the scalar framework. RD designs' credibility stems in large part from communicating the main results using simple graphs. While three-dimensional plots can communicate the intuition of two-dimensional RD designs, it is often difficult visualize model and estimation uncertainty. This makes converting boundary designs to familiar scalar designs particularly attractive.

A major choice for RD designs is how much to smooth estimates using observations away from the boundary. For two-dimensional RD designs, we derive an optimal, data-dependent, bandwidth selection rule for the conditional effect following the same general plug-in approach proposed by Imbens and Kalyanaraman (2008). Several characteristics complicate the calculation of optimal bandwidths in non-scalar settings. While local linear estimators estimators can be applied to any dimension problem, we focus primarily on two-dimensional treatment rules since the leading applications appear to take this form. Two-dimensional treatment rules include those based on two subject test scores (reading and math) and location (latitude and longitude). They also include scalar RD designs with a continuous non-forcing covariate.

We demonstrate boundary RD methods using a summer school and grade retention example studied by Jacob and Lefgren (2004) and Frandsen and Guiteras (2010). Under Chicago's accountability policy, children in third must pass both a reading and mathematics exam to avoid summer school and potentially being retained. This rule generates a two-dimensional treatment boundary. Consistent with the results in Jacob and Lefgren (2004), we find that third-grade students along the boundary who

comply with the accountability policy gain roughly 0.1 standard deviations in reading and 0.2 standard deviations in math. There is some evidence of treatment effect heterogeneity along the boundary.

The rest of the paper is organized as follows: Section 2.2 reviews scalar RD design and generalizes the notation, estimands, and identification results to boundary RD design. Section 2.3 discusses estimation of sharp and fuzzy effects by multiple local linear regression. Section 2.4 derives an optimal, data-dependent, bandwidth selection rule for conditional sharp effects. Finally, Section 2.5 provides an application and Section 2.6 concludes.

2.2 The Regression Discontinuity Model

2.2.1 Scalar RD designs

Using discontinuities in assignment rules to identify causal effects dates back to Thistlethwaite and Campbell (1960) and has a long history in psychology (Cook, 2008). However the idea has received an explosion of attention in economics after Hahn et al. (2001) formalized RD design in the language common to program evaluation. Subsequent work has further clarified RD design's underpinnings (Lee, 2008; Lee and Lemieux, 2009), developed the theory behind estimation (Porter, 2003; Sun, 2005; Lee and Card, 2008; Frandsen and Guiteras, 2010), provided practical guidance on bandwidth selection (Ludwig and Miller, 2007; Imbens and Kalyanaraman, 2008), and proposed tests of the underlying assumptions (McCrary, 2008). Imbens and Lemieux (2008) and Lee and Lemieux (2009) review this surge in RD design research, focusing on the standard scalar case.

In the standard RD setup, units have a continuous scalar covariate X and outcome Y. The treatment mechanism can be classified as "sharp" or "fuzzy" depending on whether the treatment is completely or partially determined by the covariate passing a cutoff (Trochim, 1984). In a sharp design, units receive a binary treatment $W \in \{0, 1\}$

if and only if their covariate exceeds a cutoff c. That is,

$$W = \mathbf{1}\{X \geq c\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function.

Following Neyman (1923) and Rubin (1978), we can define causal effects in terms of potential outcomes. In the potential outcomes framework, $Y(1)$ gives the outcome under treatment and $Y(0)$ gives the outcome under the control. Potential outcomes $Y(1)$ and $Y(0)$ are linked to the observed variables Y by

$$Y = Y(W) = (1 - W) \cdot Y(0) + W \cdot Y(1).$$

We only ever observe the potential outcome associated with the treatment actually received. Average causal effects are then defined as averages over unit causal effects $Y(1) - Y(0)$. Because we can only observe $Y(1)$ or $Y(0)$ but never both for a given unit, causal inference entails comparisons across potentially dissimilar units.

Intuitively, sharp regression discontinuity design identifies the average causal effect of the treatment for units at the treatment boundary ($X = c$) by comparing units ε above and below the treatment boundary as ε goes to zero. Assuming the conditional regression functions of the potential outcomes $\mathbb{E}[Y(0) \mid X = x]$ and $\mathbb{E}[Y(1) \mid X = x]$ are continuous in x, units just above and below the discontinuity have the same average potential outcomes $Y(0)$ and $Y(1)$ but differ by their treatment status W and potential outcome actually observed $Y = Y(W)$. The average causal effect of the treatment for units at the treatment boundary is identified by taking the limits from above and below,

$$\begin{aligned} \tau_{\text{SRD}} &= \mathbb{E}[Y(1) - Y(0) \mid X = c] \\ &= \mathbb{E}[Y \mid W = 1, X = c] - \mathbb{E}[Y \mid W = 0, X = c] \\ &= \lim_{x \downarrow c} \mathbb{E}[Y \mid X = x] - \lim_{x \uparrow c} \mathbb{E}[Y \mid X = x] \end{aligned} \tag{2.1}$$

where the final equality follows from continuity (Hahn et al., 2001).

The practical problem becomes how to estimate the two limits $\lim_{x \downarrow c} \mathbb{E}[Y \mid X = x]$ and $\lim_{x \uparrow c} \mathbb{E}[Y \mid X = x]$. Given the boundary nature of the problem, Hahn et al. (2001) propose estimating the limits by local linear regression. Porter (2003) proves that local linear regression is rate optimal for the regression discontinuity problem. With a rectangular kernel, local linear regression amounts to predicting both limits from a linear regression on a subset of observations around the discontinuity. Imbens and Kalyanaraman (2008) recommend local linear regression with an edge kernel and derive an optimal, data-dependent, bandwidth selection rule.

Fuzzy regression discontinuity design generalizes the RD setup to account for treatment rules that are discontinuous in the probability of treatment (Trochim, 1984; Hahn et al., 2001). That is, cases where

$$\lim_{x \downarrow c} \Pr(W = 1 \mid X = x) > \lim_{x \uparrow c} \Pr(W = 1 \mid X = x)$$

but the probabilities are not necessarily one and zero. Many real assignment rules take this form due to noncompliance with treatment assignment or individual waivers to sharp treatment rules.

As pointed out by Hahn et al. (2001), fuzzy regression discontinuity design has a strong connection to instrumental variables estimation of treatment effects with unit-varying effects (Imbens and Angrist, 1994). When a fraction of units receive treatment on each side of the cutoff, the difference between average outcomes on either side of the discontinuity becomes an intent to treat effect—the average effect of the assignment or encouragement but not the treatment. The average outcomes on both sides of the discontinuity include a mixture of treated and untreated units due to noncompliance.

To see the connection to instrumental variables, assume that units receive an encouragement or assignment Z depending on whether they fall above or below the cutoff,

$$Z = \mathbf{1}\{X \geq c\}.$$

For instance, schools may recommend students for promotion to the next grade if they score above c on an end-of-year exam but parents may choose otherwise. Given

imperfect compliance, encouragement Z may differ from treatment W. Let $W(z)$ be the treatment the unit would receive given encouragement z. Under a monotonicity assumption that there are no defiers—no units with $W(1) = 0$ and $W(0) = 1$—the fuzzy RD estimate is

$$\tau_{\text{FRD}} = \frac{\lim_{x \downarrow c} \mathbb{E}\left[Y \mid X = x\right] - \lim_{x \uparrow c} \mathbb{E}\left[Y \mid X = x\right]}{\lim_{x \downarrow c} \mathbb{E}\left[W \mid X = x\right] - \lim_{x \uparrow c} \mathbb{E}\left[W \mid X = x\right]} \tag{2.2}$$

$$= \mathbb{E}\left[Y(1) - Y(0) \mid W(1) > W(0), X = c\right],$$

which takes the form a Wald estimate (Hahn et al., 2001).

The identified parameter has the standard local average treatment effect (LATE) interpretation of the average treatment effect for units that comply with the encouragement Z and therefore have $W(1) > W(0)$. The only difference compared to a standard encouragement design setup is that the estimate is local to the encouragement cutoff c. The locality restriction arises because there is no overlap in the covariate distribution for units that receive and do not receive the encouragement and encouragement is only ignorable conditional on X. Under continuity, we can replace the exact conditioning with limits from above and below.

The practical problem is the same as for sharp RD but requires estimating the four limits in (2.2) rather than two in (2.1). Using a single bandwidth for all four conditional expectations, the local linear estimator can be implemented using weighted 2SLS with weights depending on the kernel (Imbens and Lemieux, 2008).

2.2.2 Boundary RD designs

Conceptually, boundary RD designs are similar to the scalar case except that the discontinuity cutoff becomes a boundary. In this section, we generalize the RD notation, causal parameters, and identification results to account for more general assignment rules that generate a boundary.

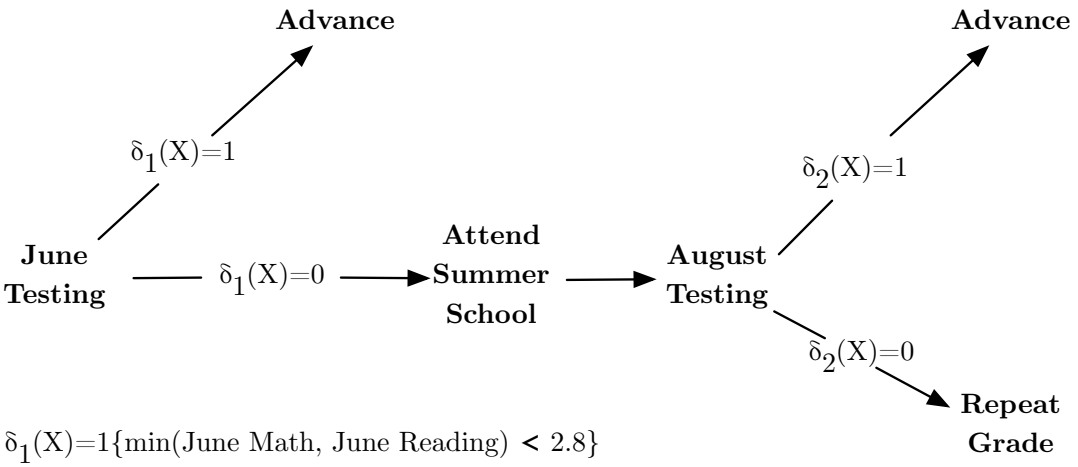


Figure 2.1: Student progress under Chicago's accountability policy.

Let $\mathbf{X}$ be a vector of d covariates. An *assignment rule,* $\delta : \boldsymbol{\mathcal{X}} \mapsto \{0, 1\}$, is a function that maps units with covariates $\mathbf{X} = \mathbf{x}$ to treatment assignment z. In the standard scalar case $d = 1$ and the assignment rule takes the simple form $\delta(x_1) = \mathbf{1}\{x_1 \geq c\}$. Our attention is on more general assignment rules that depend on a vector of covariates. For example, Chicago's summer school policy, depicted in Figures 2.1 and 2.2, assigns treatment if the minimum of two scores falls below 2.8 or $\delta(x_1, x_2) = \mathbf{1}\{\min(x_1, x_2) < 2.8\}$.

The treatment assignment rule $\delta(\mathbf{x})$ partitions the covariate space into a subset where units are assigned to treatment and where units are assigned to the control. We can therefore define the *treatment assignment set* $\mathbb{T}$ as

$$\mathbb{T} \equiv \{\mathbf{x} \in \boldsymbol{\mathcal{X}} : \delta(\mathbf{x}) = 1\}.$$

The complement of the treatment assignment set is the *control assignment set* $\mathbb{T}^c$. Using this set notation, the treatment assignment rule can be written as

$$\delta(\mathbf{x}) \equiv \mathbf{1}\{\mathbf{x} \in \mathbb{T}\}. \tag{2.3}$$

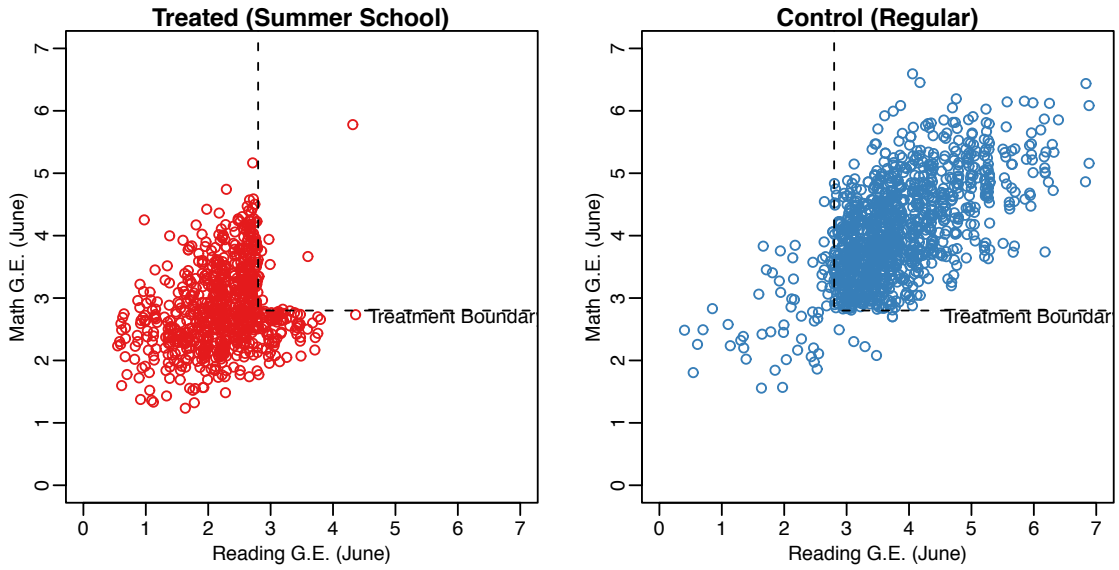


Figure 2.2: June scores and summer school attendance for a random sample of 2,000 third graders.

The treatment assignment for any unit is $Z = \delta(\mathbf{X})$. The assignment rule (2.3) is completely flexible; units receive a treatment assignment if and only if their covariates fall within an arbitrary set. For instance, Holmes (1998) estimates the effect of right-to-work laws by comparing manufacturing activity on either side of state boundaries. We could capture this identification strategy by defining $\mathbb{T}$ as the set of latitude and longitude points corresponding to right-to-work states.

With treatment assignment determined by a generic assignment rule (2.3), the discontinuity cutoff becomes a boundary. Specifically, the *assignment boundary* $\mathbb{B}$ is

$$\mathbb{B} \equiv \text{bd}(\mathbb{T}) \equiv \overline{\mathbb{T}} \cap \overline{\mathbb{T}}^c$$

where overbars denote the closure of the set. A point $\mathbf{x}$ is in the assignment boundary $\mathbb{B}$ if and only if every neighborhood around $\mathbf{x}$ contains points both in the treatment assignment set $\mathbb{T}$ and the control assignment set $\mathbb{T}^c$. For Holmes (1998) the assignment boundary consists of the borders between states that differ in right-to-work laws. For our application, the assignment boundary is illustrated in Figure 2.2.

In scalar settings, sharp RD identifies a treatment effect for units at the cutoff threshold, $X = c$. In vector settings, we can generally identify a wider set of parameters: functionals of the *conditional* treatment effect for units along the assignment boundary. To take advantage of the possibilities the assignment boundary offers, we study estimation of two related objects.

First, when all units comply with the treatment assignment, or when interest lies in the intent to treat, define the *sharp conditional treatment effect* at every point in the boundary set as

$$\tau_{\text{SBRD}}(\mathbf{x}) \equiv \mathbb{E}\left[Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}\right], \qquad \mathbf{x} \in \mathbb{B}. \tag{2.4}$$

This parameter illustrate the advantage of having multiple dimensions: the ability to estimate a treatment effect along the entire boundary rather than a single point. Second, define the *sharp average treatment effect* as

$$\tau_{\text{SBRD}} \equiv \mathbb{E}\left[Y(1) - Y(0) \mid \mathbf{X} \in \mathbb{B}\right] \tag{2.5}$$

by taking the expectation over the boundary. The average treatment effect yields a single summary measure of the treatment effect along the boundary. In Holmes's (1998) application, the conditional sharp effect $\tau_{\text{SBRD}}(\mathbf{x})$ measures the impact of right-to-work laws at any geographic point along the boundaries between states with different right-to-work laws. The average effect τ_{SBRD} measures the average effect along the boundary.

For RD designs, identification arises by looking at observations near the boundary. To formalize identification in higher dimensions, denote the ϵ-neighborhood around $\mathbf{x}$ as $N_\epsilon(\mathbf{x})$. The ϵ-neighborhood around $\mathbf{x}$ contains all points $\mathbf{X}$ within a sphere of radius ϵ around $\mathbf{x}$. That is, $N_\epsilon(\mathbf{x}) \equiv \{\mathbf{X} \in \boldsymbol{\mathcal{X}} : (\mathbf{X} - \mathbf{x})'(\mathbf{X} - \mathbf{x}) < \epsilon^2\}$. Further, let $N_\epsilon^+(\mathbf{x}) \equiv N_\epsilon(\mathbf{x}) \cap \mathbb{T}$ be the points in ϵ-neighborhood around $\mathbf{x}$ that receive treatment and $N_\epsilon^-(\mathbf{x}) \equiv N_\epsilon(\mathbf{x}) \cap \mathbb{T}^c$ be the points the receive the control. For average effects, it is convenient to define a neighborhood of the boundary *set* and let this neighborhood contract. Let $B_\epsilon \equiv \{\mathbf{X} \in \boldsymbol{\mathcal{X}} : \exists \mathbf{x} \in \mathbb{B} \text{ s.t. } \mathbf{X} \in N_\epsilon(\mathbf{x})\}$ be points $\mathbf{X}$ within an ϵ-neighborhood of any point $\mathbf{x}$ in the boundary $\mathbb{B}$. Similarly, define $B_\epsilon^+ \equiv B_\epsilon \cap \mathbb{T}$ and

$B_\epsilon^- \equiv B_\epsilon \cap \mathbb{T}^c$ as the set of associated treatment and control points. Frölich (2007) using similar notation to study regression discontinuity design with covariates.

To identify the sharp effects, we make two assumptions. First:

Assumption 2.2.1. (Boundary Positivity) *For all* $\mathbf{x} \in \mathbb{B}$ *and* $\epsilon > 0$, $\Pr(\mathbf{X} \in N_\epsilon^-(\mathbf{x})) > 0$ *and* $\Pr(\mathbf{X} \in N_\epsilon^+(\mathbf{x})) > 0$.

The boundary positivity assumption ensures that both treated and untreated units exist along the boundary. Second:

Assumption 2.2.2. (Continuity) *The conditional regression functions* $\mathbb{E}[Y(1) \mid \mathbf{X} = \mathbf{x}]$ *and* $\mathbb{E}[Y(0) \mid \mathbf{X} = \mathbf{x}]$ *are continuous in* $\mathbf{x}$. *Further, the marginal density* $f_{\mathbf{X}}$ *is continuous in* $\mathbf{x}$.

Continuity of the conditional regression functions ensures that units in the neighborhood of the boundary have comparable potential outcomes. We could weaken this assumptions to require continuity only near the boundary set $\mathbb{B}$, however it is difficult to imagine instances where one would hold without the other. While not required for identification, continuity of the density ensures average effects do not depend on which side of the boundary the integration is taken over. Combined, positivity and continuity yield the standard RD identification result (Hahn et al., 2001) but for boundary RD designs:

Theorem 2.2.3. (Sharp Boundary RD) *Under Assumption 2.2.1 and 2.2.2, for all* $\mathbf{x} \in \mathbb{B}$,

$$\tau_{SBRD}(\mathbf{x}) = \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_\epsilon^+(\mathbf{x})\right] - \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_\epsilon^-(\mathbf{x})\right]. \tag{2.6}$$

$$\tau_{SBRD} = \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in B_\epsilon^+\right] - \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in B_\epsilon^-\right]. \tag{2.7}$$

Proof. For all $\mathbf{x} \in \mathbb{B}$,

$$\begin{aligned}
\tau_{\text{SBRD}}(\mathbf{x}) &= \mathbb{E}\left[Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}\right] \\
&= \lim_{\varepsilon \to 0} \mathbb{E}\left[Y(1) \mid \mathbf{X} \in N_\epsilon^+(\mathbf{x})\right] - \lim_{\varepsilon \to 0} \mathbb{E}\left[Y(0) \mid \mathbf{X} \in N_\epsilon^-(\mathbf{x})\right] \\
&= \lim_{\varepsilon \to 0} \mathbb{E}\left[Y(1) \mid \mathbf{X} \in N_\epsilon^+(\mathbf{x}), W = 1\right] - \lim_{\varepsilon \to 0} \mathbb{E}\left[Y(0) \mid \mathbf{X} \in N_\epsilon^-(\mathbf{x}), W = 0\right] \\
&= \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_\epsilon^+(\mathbf{x})\right] - \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_\epsilon^-(\mathbf{x})\right].
\end{aligned}$$

The average RD parameter τ_{VSRD} follows immediately by taking expectation over $\mathbf{x} \in \mathbb{B}$ or by the same argument with a different conditioning set. □

In many applications, units do not perfectly comply with the assigned treatment. For instance, some students receive discretionary waivers under the Chicago accountability policy we study. As in the scalar case, we can define a fuzzy effect following the instrumental variables intuition. The *fuzzy conditional treatment effect* is

$$\tau_{\text{FBRD}}(\mathbf{x}) \equiv \mathbb{E}\left[Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}, W(1) > W(0)\right], \qquad \mathbf{x} \in \mathbb{B}. \tag{2.8}$$

The fuzzy effect measures the treatment's impact on units that comply with the assignment Z, $W(1) > W(0)$, and have covariates $\mathbf{X} = \mathbf{x}$. Again, the average treatment effect over the entire assignment boundary defines the *fuzzy average treatment effect*

$$\tau_{\text{FBRD}} \equiv \mathbb{E}\left[Y(1) - Y(0) \mid \mathbf{X} \in \mathbb{B}, W(1) > W(0)\right]. \tag{2.9}$$

When compliance with the treatment is imperfect, we make the standard instrumental variables first-stage and monotonicity assumptions but applied to RD. The first-stage assumption ensures that units assigned to treatment and control exist in the neighborhood of the boundary and that a boundary discontinuity exists in the probability of treatment.

Assumption 2.2.4. (First Stage) *For all* $\mathbf{x} \in \mathbb{B}$,

$$\lim_{\varepsilon \to 0} \mathbb{E}\left[W \mid \mathbf{X} \in N_{\epsilon}^{+}(\mathbf{x})\right] > \lim_{\varepsilon \to 0} \mathbb{E}\left[W \mid \mathbf{X} \in N_{\epsilon}^{-}(\mathbf{x})\right].$$

We also impose a monotonicity or no defiers assumption that rules out units that take the control when assigned to the treatment and take the treatment when assigned to the control. For Chicago's accountability policy, monotonicity implies no units would attend summer school if they passed the end-of-year exam but would not attend summer school if they failed.

Assumption 2.2.5. (Monotonicity/No Defiers) $W(1) \geq W(0)$ for all units.

Finally, we assume that the conditional regression functions for compliance is continuous.

Assumption 2.2.6. (Continuity of Compliance) *The conditional regression functions* $\mathbb{E}[W(1) \mid \mathbf{X} = \mathbf{x}]$ *and* $\mathbb{E}[W(0) \mid \mathbf{X} = \mathbf{x}]$ *are continuous in* $\mathbf{x}$.

Theorem 2.2.7 gives the key identification result for conditional fuzzy RD:

Theorem 2.2.7. *Under Assumption 2.2.1, 2.2.2, 2.2.4, 2.2.5, and 2.2.6, for all* $\mathbf{x} \in \mathbb{B}$

$$\tau_{FBRD}(\mathbf{x}) = \lim_{\varepsilon \to 0} \frac{\mathbb{E}[Y \mid \mathbf{X} \in N_{\epsilon}^{+}(\mathbf{x})] - \mathbb{E}[Y \mid \mathbf{X} \in N_{\epsilon}^{-}(\mathbf{x})]}{\mathbb{E}[W \mid \mathbf{X} \in N_{\epsilon}^{+}(\mathbf{x})] - \mathbb{E}[W \mid \mathbf{X} \in N_{\epsilon}^{-}(\mathbf{x})]}, \tag{2.10}$$

$$\tau_{FBRD} = \lim_{\varepsilon \to 0} \frac{\mathbb{E}[Y \mid \mathbf{X} \in B_{\epsilon}^{+}] - \mathbb{E}[Y \mid \mathbf{X} \in B_{\epsilon}^{-}]}{\mathbb{E}[W \mid \mathbf{X} \in B_{\epsilon}^{+}] - \mathbb{E}[W \mid \mathbf{X} \in B_{\epsilon}^{-}]}. \tag{2.11}$$

Proof. The proof for the conditional effect follows directly from Hahn et al. (2001) Theorem 3 but with alternate notation and generalized to vector $\mathbf{x}$. The average effect follows the same argument but with an different conditioning set. See appendix for details. □

For the conditional effect, Theorem 2.2.7 is simply the standard instrumental variables analogy laid out by Hahn et al. (2001) for scalar fuzzy regression discontinuity

design. However, rather than defining a limit from above and below we instead consider an ε-neighborhood around the boundary point $\mathbf{x}$ and let $\varepsilon \to 0$. The "above" and "below" that was central to the identification in the scalar case effectively comes from conditioning on the assignment Z. Conditional on $Z = 1$, or, equivalently, $\mathbf{X} \in \mathbb{T}$, all observations in the ε-neighborhood around $\mathbf{x}$ receive treatment $W(1)$ and have outcome $Y = Y(0) \cdot (1 - W(1)) + Y(1) \cdot W(1)$. The average effect looks at points near the boundary rather than a single point.

2.3 Estimation

2.3.1 Local linear estimation of conditional effects

2.3.1.1 Sharp conditional effect

The sharp conditional effect is a simple functional of the two limits

$$m_0(\mathbf{x}) = \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_\epsilon^-(\mathbf{x})\right],$$

and

$$m_1(\mathbf{x}) = \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_\epsilon^+(\mathbf{x})\right].$$

Specifically, $\tau_{\text{SBRD}}(\mathbf{x}) = m_1(\mathbf{x}) - m_0(\mathbf{x})$. Estimation of the two limits $m_0(\mathbf{x})$ and $m_1(\mathbf{x})$ is a standard nonparametric regression problem at the boundary. As such, estimation of the conditional RD parameter $\tau_{\text{SBRD}}(\mathbf{x})$ is not fundamentally different than when $\mathbf{x}$ is scalar. For the scalar case, Hahn et al. (2001) suggest estimating the limits using local linear regression. Local linear regression has appealing boundary properties (Fan, 1992; Fan and Gijbels, 1996) and is rate optimal for the regression discontinuity problem (Porter, 2003). These optimality properties extend to boundary RD design. Ruppert and Wand (1994) study the multivariate version of local linear regression that we apply.

Consider estimation of $m_0(\mathbf{x})$. Let $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$ be a d-variate kernel with positive definite bandwidth matrix $\mathbf{H}^{1/2}$. The local linear estimator for $m_0(\mathbf{x})$ solves

$$\min_{\alpha_0,\boldsymbol{\beta}_0} \sum_{i\in\{i:W_i=0\}} K_{\mathbf{H}_0}(\mathbf{X}_i - \mathbf{x}) \cdot \left(Y_i - \alpha_0 + (\mathbf{X}_i - \mathbf{x})'\boldsymbol{\beta}_0\right)^2$$

and returns $\hat{\alpha}_0$ as the estimate for $m_0(\mathbf{x})$. The solution takes the form of a weighted least squares estimate

$$\hat{m}_0(\mathbf{x}) = \hat{\alpha}_0 = \mathbf{e}_1' \left(\sum_{\{i:W_i=0\}} \mathbf{V}_i b_i \mathbf{V}_i' \right)^{-1} \sum_{\{i:W_i=0\}} \mathbf{V}_i b_i Y_i, \tag{2.12}$$

where $\mathbf{V}_i = [\ 1 \quad (\mathbf{X}_i - \mathbf{x})'\ ]'$, $\mathbf{e}_1$ is a $(d+1)\times 1$ vector of zeros with the first element equal to one, and b_i are weights equal to $K_{\mathbf{H}_0}(\mathbf{X}_i - \mathbf{x})$. The limit $m_1(\mathbf{x})$ can be estimated following (2.12) but with the summation over treated units $W_i = 1$ and bandwidth matrix $\mathbf{H}_1^{1/2}$. A plug-in estimator for the sharp conditional effect is then $\hat{\tau}_{\text{VSRD}}(\mathbf{x}) = \hat{m}_1(\mathbf{x}) - \hat{m}_0(\mathbf{x})$. A plug-in estimator for the estimator's variance can be calculated using the estimated residuals in the same manner as Imbens and Lemieux (2008) give for the scalar case.

While the plug-in approach is relatively simple, it is computationally convenient to frame the conditional sharp RD effect as ordinary least squares (OLS). Let

$$\mathbf{R}_i = \begin{bmatrix} 1 \\ W_i \\ Z_i \cdot (\mathbf{X}_i - \mathbf{x}) \\ (1 - Z_i) \cdot (\mathbf{X}_i - \mathbf{x}) \end{bmatrix} \tag{2.13}$$

be a vector of regressors for unit i, where $Z_i = W_i$ in the sharp RD case considered here. Then the local linear estimate for $\tau_{\text{SBRD}}(\mathbf{x})$ can be written in terms of weighted OLS as

$$\hat{\tau}_{\text{SBRD}}(\mathbf{x}) = \mathbf{e}_2' \left(\mathbf{R}'\mathbf{B}\mathbf{R}\right)^{-1} \mathbf{R}'\mathbf{B}\mathbf{Y}, \tag{2.14}$$

where $\mathbf{e}_2$ is $(2d+2) \times 1$ vector of zeros with the second element equal to one and $\mathbf{B}$ is an $n \times n$ weight matrix with diagonal elements equal to $K_{\mathbf{H}_1}(\mathbf{X}_i - \mathbf{x})$ for units with $Z_i = 1$ and $K_{\mathbf{H}_0}(\mathbf{X}_i - \mathbf{x})$ for units with $Z_i = 0$. Intuitively, the estimator (2.14) fits two weighted regressions with differing slopes on either side of the boundary. Given the re-centering $(\mathbf{X}_i - \mathbf{x})$, the coefficient on the second covariate W_i provides the local linear estimate for the discontinuous increase or decrease in the outcome Y at the point $\mathbf{x}$. The estimator (2.14) can easily be implemented in any major statistical package. Assuming the bandwidth matrices $\mathbf{H}_1$ and $\mathbf{H}_0$ are selected such that the asymptotic bias disappears, valid inference is based on the robust OLS standard errors. That is, $\hat{\mathbb{V}}\left(\hat{\tau}_{\text{VFRD}}(\mathbf{x}) \mid \mathbf{R}\right) = \mathbf{P}\hat{\Omega}\mathbf{P}'$ where $\mathbf{P} = \mathbf{e}_2' \left(\mathbf{R}'\mathbf{B}\mathbf{R}\right)^{-1} \mathbf{R}'\mathbf{B}$ and $\hat{\Omega}$ is a diagonal matrix of squared residuals $\hat{\varepsilon} = \mathbf{Y} - \mathbf{R}\left(\mathbf{R}'\mathbf{B}\mathbf{R}\right)^{-1} \mathbf{R}'\mathbf{B}\mathbf{Y}$.

The estimator (2.14) requires a bandwidth matrix for treated and untreated units, $\mathbf{H}_1$ and $\mathbf{H}_0$. We discuss the choice of bandwidth matrices in Section 2.4. In general, is is usually adequate to use a simple kernel with equivalent bandwidths matrices for $\mathbf{H}_1$ and $\mathbf{H}_0$. However, the weighted OLS perspective does not enforce restrictive choices for the bandwidth matrices or kernel.

2.3.1.2 Fuzzy conditional effect

The fuzzy conditional effect $\tau_{\text{FBRD}}(\mathbf{x})$ is a functional of the four limits

$$n_0(\mathbf{x}) = \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_\epsilon^-(\mathbf{x})\right],$$

$$n_1(\mathbf{x}) = \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_\epsilon^+(\mathbf{x})\right],$$

$$p_0(\mathbf{x}) = \lim_{\varepsilon \to 0} \mathbb{E}\left[W \mid \mathbf{X} \in N_\epsilon^-(\mathbf{x})\right],$$

and

$$p_1(\mathbf{x}) = \lim_{\varepsilon \to 0} \mathbb{E}\left[W \mid \mathbf{X} \in N_\epsilon^+(\mathbf{x})\right].$$

Specifically,

$$\tau_{\text{FBRD}}(\mathbf{x}) = \frac{n_1(\mathbf{x}) - n_0(\mathbf{x})}{p_1(\mathbf{x}) - p_0(\mathbf{x})}. \tag{2.15}$$

An intuitive strategy would be to estimate the limits using four local linear regression with bandwidth matrices $\mathbf{H}_{n_1}$, $\mathbf{H}_{n_1}$, $\mathbf{H}_{p_1}$ and $\mathbf{H}_{p_1}$. Plugging these estimates into (2.15) would give an estimate of the fuzzy conditional effect. However, following the instrumental variables intuition, it is convenient to frame fuzzy RD as two-stage least squares (TSLS). Let $\mathbf{S}$ be an $n \times (2d+2)$ matrix of instruments with rows equal to

$$\mathbf{S}_i = \begin{bmatrix} 1 \\ Z_i \\ Z_i \cdot (\mathbf{X}_i - \mathbf{x}) \\ (1 - Z_i) \cdot (\mathbf{X}_i - \mathbf{x}) \end{bmatrix} \tag{2.16}$$

and $\mathbf{R}_i$ be defined as in (2.13), now with $Z_i \neq W_i$ for some i. Thus, Z_i in $\mathbf{S}_i$ is the excluded instrument and W_i in $\mathbf{R}_i$ is the endogenous variable. The weighted TSLS estimator for $\tau_{\text{VFRD}}(\mathbf{x})$ is then

$$\hat{\tau}_{\text{FBRD}}(\mathbf{x}) = \mathbf{e}_2' \left(\mathbf{R}'\mathbf{B}_2\mathbf{S}(\mathbf{S}'\mathbf{B}_1\mathbf{S})^{-1}\mathbf{S}'\mathbf{B}_1\mathbf{R}\right)^{-1} \mathbf{R}'\mathbf{B}_2\mathbf{S}(\mathbf{S}'\mathbf{B}_1\mathbf{S})^{-1}\mathbf{S}'\mathbf{B}_1\mathbf{Y} \tag{2.17}$$

where $\mathbf{e}_2$ is $(2d+2) \times 1$ vector of zeros with the second element equal to one, and $\mathbf{B}_1$ and $\mathbf{B}_2$ are $n \times n$ weight matrices for the first and second stages. To make the estimator (2.17) equivalent to the plug-in estimator, the weight matrices would depend on all four bandwidth matrices. Specifically, the first stage $\mathbf{B}_1$ would have diagonal elements $K_{\mathbf{H}_{p0}}(\mathbf{X}_i - \mathbf{x})$ and $K_{\mathbf{H}_{p1}}(\mathbf{X}_i - \mathbf{x})$ and the second stage $\mathbf{B}_2$ would have diagonal elements $K_{\mathbf{H}_{n0}}(\mathbf{X}_i - \mathbf{x})$ and $K_{\mathbf{H}_{n1}}(\mathbf{X}_i - \mathbf{x})$ for units with $Z_i = 0$ and $Z_i = 1$, respectively. In general, defining a single bandwidth matrix for both the first and second stage and units assigned to the treatment and control is usually adequate. Section 2.4 discusses bandwidth and kernel selection in detail.

The estimate of the conditional fuzzy RD parameter $\hat{\tau}_{\text{FBRD}}(\mathbf{x})$ given in (2.14) is a weighted TSLS estimate of the coefficient on the endogenous variable W_i with an excluded instrument Z_i. When the discontinuity is sharp and $Z_i = W_i$ the prediction reduces to weighted OLS.

Assuming suitable under-smoothing, inference follows from the robust TSLS standard errors. That is,

$$\hat{\mathbb{V}}\left(\hat{\tau}_{\text{VFRD}}(\mathbf{x}) \mid \mathbf{R}, \mathbf{S}\right) = \mathbf{Q}\hat{\Omega}\mathbf{Q}'$$

where

$$\mathbf{Q} = \mathbf{e}_2' \left(\mathbf{R}'\mathbf{B}_2\mathbf{S}(\mathbf{S}'\mathbf{B}_1\mathbf{S})^{-1}\mathbf{S}'\mathbf{B}_1\mathbf{R}\right)^{-1} \mathbf{R}'\mathbf{B}_2\mathbf{S}(\mathbf{S}'\mathbf{B}_1\mathbf{S})^{-1}\mathbf{S}'\mathbf{B}_1$$

and $\hat{\Omega}$ is a diagonal matrix of squared TSLS residuals. This is a weighted version of the standard robust variance estimator for TSLS. The robust TSLS variance is numerically identical to the plug-in estimator. When the discontinuity is sharp, the robust TSLS variance reduces to the robust OLS variance. Both can be straightforwardly estimated using standard statistical packages provided $\mathbf{B}_1 = \mathbf{B}_2$.

2.3.2 Estimation of average effects

A natural approach to estimating the average effects, as written in Theorems 2.2.3 and 2.2.7, is to take points near the boundary as estimates of the associated limits. In finite samples, however, adjusting for how far points are from the boundary can reduce bias. This has led many researchers, particularly in geographic applications, to estimate average effects by reducing the two-dimensional RD problem to scalar RD with "distance to the nearest boundary" as the univariate forcing variable (e.g., Holmes, 1998; Black, 1999; Kane et al., 2006; Davidoff and Leigh, 2008; Lalive, 2008; Dell, 2010).

Regression discontinuity designs with "distance to the nearest boundary" as the univariate forcing variable can be estimated by the methods reviewed in Imbens and Lemieux (2008) and Lee and Lemieux (2009). Scalar RD consistently estimates both the sharp and average effects, and the scalar setup simplifies the graphical presentation. Bandwidth selection options include cross-validation methods proposed by Ludwig and Miller (2007) and plug-in rules proposed by Imbens and Kalyanaraman (2008).

An alternative to reducing the problem to scalar RD is to integrate the conditional effects explicitly over the boundary. For instance, the sharp average effect can be written as

$$\tau_{\text{SBRD}} = \int_{\mathbf{x}\in\mathbb{B}} \tau_{\text{SBRD}}(\mathbf{x}) f(\mathbf{x} \mid \mathbf{X} \in \mathbb{B}) d\mathbf{x} = \frac{\int_{\mathbf{x}\in\mathbb{B}} \tau_{\text{SBRD}}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}}{\int_{\mathbf{x}\in\mathbb{B}} f(\mathbf{x}) d\mathbf{x}}. \tag{2.18}$$

This path integral suggests estimating the conditional effect and density at each point along the boundary and integrating explicitly. Combining multiple local linear regressions for the conditional effects and kernel density estimates for the density will generally have better finite sample properties than just using distance to the boundary in one dimension.

A simple approach to numerically integrate over the boundary is a fixed design. Choose K evenly spaced points $\mathbf{x}_k$ along the boundary, indexed by $k = 1, \ldots, K$.

Then, estimate the average sharp effect as

$$\hat{\tau}_{\text{SBRD}} = \frac{\sum_{k=1}^{K} \hat{\tau}_{\text{SBRD}}(\mathbf{x}_k) \cdot \hat{f}(\mathbf{x}_k)}{\sum_{k=1}^{K} \hat{f}(\mathbf{x}_k)}. \tag{2.19}$$

As K increases, the summation over (2.19) will approach the integrals in (2.18). In most applications, a modest K will suffice provided that the conditional effects and density are smooth. A good rule-of-thumb is to choose K such that the boundary is reasonably well covered and then check that the estimated effect and standard errors do not change if K doubles.

After explicitly integrating over the boundary, a final step is to derive a consistent estimate for the variance of average effect $\hat{\tau}_{\text{VSRD}}$. One possibility would be to use resampling methods. In Appendix 2.A.2, we derive the influence function for the average effect estimated using a kernel density and multiple local linear regression and give a variance estimate based the nonparametric delta method (Huber, 1981; Newey, 1994; Davison and Hinkley, 1997). Intuitively, the variance depends both on the uncertainty involved in estimating the density and conditional expectation.

The average fuzzy effect, τ_{FBRD}, requires integration over the density of $\mathbf{X}$ for compliers $f(\mathbf{x} \mid W(1) > W(0), \mathbf{X} \in \mathbb{B})$ not the density for all units $f(\mathbf{x} \mid \mathbf{X} \in \mathbb{B})$. Abadie (2003) shows that all functions of the joint distribution of $(Y, Z, \mathbf{X})$, including the marginal density of $\mathbf{X}$, are identified for compliers. We can therefore estimate the average effect as

$$\hat{\tau}_{\text{FBRD}} = \frac{\sum_{k=1}^{K} \hat{\tau}_{\text{FBRD}}(\mathbf{x}_k) \cdot \hat{\lambda}(\mathbf{x})}{\sum_{k=1}^{K} \hat{\lambda}(\mathbf{x})} \tag{2.20}$$

where $\lambda(\mathbf{x})$ is defined by (2.23) in Appendix 2.A.1 and expresses $f(\mathbf{x} \mid W(1) > W(0), \mathbf{X} \in \mathbb{B})$ in terms of the observed joint distribution of $(Y, Z, \mathbf{X})$. A influence

function based variance estimate could be calculated similarly to the sharp effect, although with additional terms.

We expect that explicitly integrating over the boundary will rarely differ from reducing the problem to one dimension and using distance to the nearest boundary as a scalar forcing variable. Given the additional complexity involved in integrating explicitly, we recommend estimating average effects using scalar RD methods reviewed in Imbens and Lemieux (2008) and Lee and Lemieux (2009). By using distance to the nearest boundary the scalar forcing variable binds along the entire boundary.

2.3.3 Graphical approaches to boundary RD designs

A major appeal of RD is the ability to visualize the identification strategy and resulting estimate. It is now standard practice report both the specific point estimates and give visual evidence of the discontinuity in treatment and outcomes. In higher dimensions, equivalent graphs can be difficult to construct. With two forcing variables, such as reading and math scores, the boundary discontinuity can be visualized using three-dimensional or contour plots. However, in our experience, it is difficult to construct such graphs that are informative about inference. For instance, a visual break near the discontinuity almost always occurs even if it is not statistically significant, due to the boundary nature of the estimation. But model and estimation uncertainty is hard to report graphically in three dimensions.

An alternative approach is to base the visual RD evidence on the average effect. The average effect can be estimated consistently using distance to the nearest boundary as a scalar forcing variable. This suggest using the standard RD graphs (e.g., Imbens and Lemieux, 2008) with distance to the nearest boundary as the forcing variable. Specifically, we suggest plotting the average outcome or treatment probability and confidence intervals for under-smoothed bins of distance on each side of the discontinuity, along with the local linear or series regression fitted values and confidence intervals. Many researchers in geographic applications have used this strategy successfully. The downside of this approach is that the average effect may mask discontinuities at certain points along the boundary. However, if interest lies primarily in

the average effect, reducing boundary RD designs into the scalar framework enables the standard RD graphs that are central to RD designs' appeal.

2.4 Bandwidth Selection

A major remaining issue is the choice of bandwidth matrices. We consider this choice for two-dimensional boundary RD designs. Following Imbens and Kalyanaraman (2008), we derive a plug-in bandwidth selection rule that minimizes the asymptotic conditional mean square error (AMSE) of the sharp RD estimate. The conditional MSE for the sharp estimates is $\mathbb{E}\left[\left(\hat{\tau}_{\text{SBRD}}(x_1, x_2) - \tau_{\text{SBRD}}(x_1, x_2)\right)^2 \mid \mathbf{X}\right]$, which we take as the objective function. In most fuzzy RD applications it is more difficult to estimate the jump in outcome rather than jump in treatment probability. An optimal bandwidth for the sharp effect is therefore likely to perform well for the fuzzy effect as well.

Under regularity conditions given in Appendix 2.A.3, Ruppert and Wand (1994) give the conditional asymptotic bias and variance for multiple local linear regression estimates of boundary points such as $m_0(\mathbf{x})$ and $m_1(\mathbf{x})$ where $\mathbf{x} \in \mathbb{B}$. The conditional asymptotic bias and variance can be combined to form an estimate of the AMSE of $\tau_{\text{SBRD}}(\mathbf{x})$.

There are a number of practical difficulties in this approach. First, in general, there is no closed form solution for the optimal choice of the bandwidth matrix $\mathbf{H}^{1/2}$. Second, for points on the boundary, the optimal bandwidth depends the boundary's precise shape. In scalar RD, the kernel is always cut in half by the boundary, which simplifies calculations. In higher dimensions, however, the treatment assignment boundary can cut through the kernel in an arbitrary fashion. Third, the optimal bandwidth requires estimates of the Hessians of $m_0(\mathbf{x})$ and $m_1(\mathbf{x})$. If estimated by local polynomial regression, these estimates require pilot bandwidths. Finally, the optimal bandwidth expression suffers from two instabilities. As in Imbens and Kalyanaraman (2008), the the first-order asymptotic biases of $m_0(\mathbf{x})$ and $m_1(\mathbf{x})$ cancel when the Hessians take specific forms, leading to an infinite optimal bandwidth. Unlike Imbens and Kalyanaraman (2008), the asymptotic bias of $m_0(\mathbf{x})$ and $m_1(\mathbf{x})$

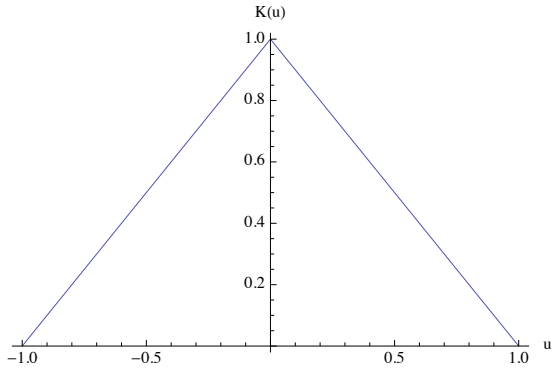
can also both equal zero independently in multivariate problems. Moreover, these bias cancellations depend both on the Hessian and the precise shape of the boundary, implying the shape the boundary can lead to an arbitrary large change in the optimal bandwidth.

To address these difficulties, we make a number of simplifying assumptions. First, we consider a two-dimensional problem with diagonal bandwidth matrix $\mathbf{H}^{1/2} = \text{diag}([\ \sigma_1 h \quad \sigma_2 h\ ])$, where σ_1 and σ_2 are the standard deviations for each dimensions and h is a common scalar bandwidth. While the choice of a single bandwidth h significantly limits the kernel's flexibility, we have found it performs reasonably if the forcing variables are normalized onto the same scale. Second, we use a two-dimensional edge product kernel

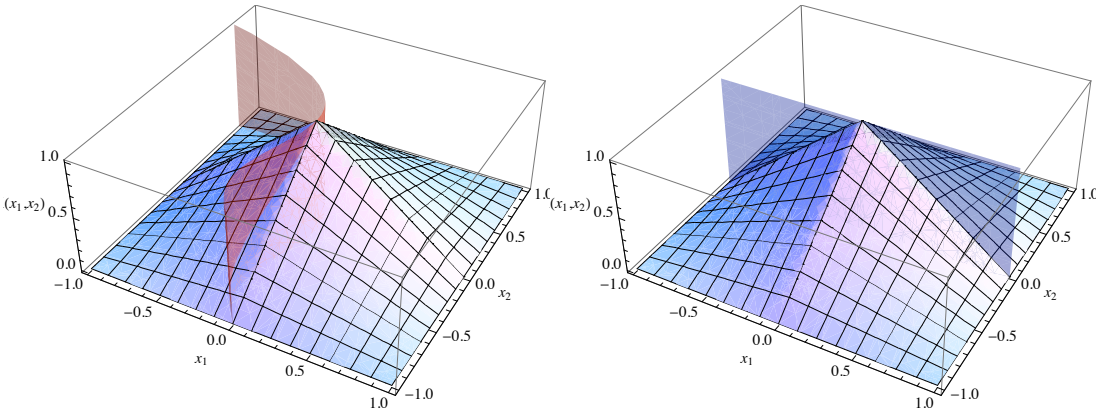
$$K(u_1, u_2) = (1 - |u_1|) \cdot (1 - |u_2|) \cdot \mathbf{1}\{|u_2| \leq 1, |u_2| \leq 1\},$$

which has boundary optimality properties in the scalar case (Cheng et al., 1997). Third, we assume the conditional variance function $\upsilon(\mathbf{x}) = \mathbb{V}(Y_i \mid \mathbf{X}_i = \mathbf{x})$ and the density $f(\mathbf{x})$ is continuous across the boundary and are bounded away from zero over the support of $\mathbb{B}$.

Finally, the primary difficulty in multivariate settings is the interaction between the the boundary and the kernel. As shown in Appendix 2.A.3, the conditional AMSE depends on how the boundary cuts through the kernel at any given point. Figure 2.3 illustrates this phenomenon. To circumvent the need to calculate an AMSE expression for every type of boundary, we consider two special cases: vertical and horizontal boundaries in two dimensions estimated at points away from vertices. For other types of boundary points, we recommend choosing one of these two special types as an approximation.



(a) One-dimensional edge kernel.



(b) Two-dimensional edge kernel.

Figure 2.3: Interaction between treatment boundary and kernel. In two dimensions, the volume under the kernel on each side of the boundary depends on the boundary's precise shape.

Given regularity conditions and assumptions stated in Appendix 2.A.3, the AMSE of $\tau_{\text{SBRD}}(x_1, x_2)$ for such points is

$$\text{AMSE}(h; \mathbf{x}) = \frac{h^4}{3600} \cdot \left(C_1 \cdot \sigma_1^2 \cdot \left[m_0^{11}(\mathbf{x}) - m_1^{11}(\mathbf{x}) \right] + C_2 \cdot \sigma_2^2 \cdot \left[m_1^{22}(\mathbf{x}) - m_0^{22}(\mathbf{x}) \right] \right)^2 + \frac{32 \cdot \upsilon(\mathbf{x})}{5 \cdot f(\mathbf{x}) \cdot h^2 \cdot \sigma_1 \cdot \sigma_2 \cdot n} + o_p\{\frac{1}{nh^2} + h^4\},$$

where $C_1 = 5$ and $C_2 = 3$ for points on a horizontal boundary and $C_1 = 3$ and $C_2 = 5$ for points on a vertical boundary; and $m_w^{11}(x_1, x_2)$ and $m_w^{22}(x_1, x_2)$ denote the diagonal elements of the Hessian for $m_w(x_1, x_2)$ with $w = 0, 1$.

Intuitively, the squared bias component of the AMSE depends on the curvature of the conditional regression functions on either side of the boundary and the variance component of the AMSE depends on the conditional variance, density, and bandwidth. The constants C_1 and C_2 in the bias term arise from how we assume the boundary cuts through the kernel. For horizontal and vertical boundaries, the cross-derivative terms in the Hessians to not appear in the AMSE.

Minimizing the AMSE by differentiating and solving for h yields the optimal plug-in bandwidth for $\tau_{\text{VSRD}}(x_1, x_2)$

$$h_{\text{opt}}(\mathbf{x}) \approx 4.75 \cdot \left(\frac{\upsilon(\mathbf{x}) / (f(\mathbf{x}) \cdot \sigma_1 \cdot \sigma_2)}{\left(C_1 \cdot \sigma_1^2 \cdot [m_0^{11}(\mathbf{x}) - m_1^{11}(\mathbf{x})] + C_2 \cdot \sigma_2^2 \cdot [m_1^{22}(\mathbf{x}) - m_0^{22}(\mathbf{x})] \right)^2} \right)^{1/6} \cdot n^{-1/6}. \tag{2.21}$$

There are eight unknowns: the standard deviations of the forcing variables σ_1 and σ_2, the conditional outcome variance $\upsilon(\mathbf{x})$, the density $f(\mathbf{x})$, and the four second derivatives $m_0^{11}(\mathbf{x})$, $m_1^{11}(\mathbf{x})$, $m_0^{22}(\mathbf{x})$, and $m_1^{22}(\mathbf{x})$.

The optimal bandwidth is potentially unstable because the denominator may be close to zero. This instability arises because the first-order asymptotic biases can

cancel, leaving only the asymptotic variance term that is minimized by an infinite bandwidth. Unlike Imbens and Kalyanaraman (2008), the bias cancellation is not unique to the RD setup and also arises because of the multivariate nature of the estimation. Imbens and Kalyanaraman (2008) and Kalyanaraman (2009) propose regularization strategies to overcome ill-posedness in bandwidth selection problems. Here we suggest an alternative, simpler, strategy.

Unlike the scalar RD case for which there is a single optimal bandwidth, the optimal bandwidth (2.21) varies along the boundary. Instead of formal regularization, we suggest calculating $h_{\text{opt}}(\mathbf{x})$ for a set of evenly spaced points along the boundary and then taking the minimum as the final bandwidth choice for each direction. Selecting the minimum bandwidth will also generally ensure a suitable degree of under-smoothing that justifies robust standard errors.

Following this strategy, a rule-of-thumb bandwidth selection rule can be derived by plugging in estimated values for the unknown quantities in (2.21) for each candidate point and then taking the minimum bandwidths as the final choice. More precisely, we

1. Estimate the standard deviations σ_1 and σ_2, conditional variance $\upsilon(\mathbf{x})$, and density $f(\mathbf{x})$.

2. Estimate the second derivatives $m_0^{11}(\mathbf{x})$, $m_1^{11}(\mathbf{x})$, $m_0^{22}(\mathbf{x})$, and $m_1^{22}(\mathbf{x})$.

3. Calculate $\hat{h}_{\text{opt}}(\mathbf{x}_k)$ for K evenly spaced points.

4. Select the minimum $\hat{h}_{\text{opt}}(\mathbf{x}_k)$ as the rule-of-thumb bandwidth h_{ROT}.

We now describe each step in more detail.

Step 1: Estimate the standard deviations σ_1 and σ_2, conditional variance $\upsilon(\mathbf{x})$, and density $f(\mathbf{x})$:

First, calculate the standard deviations of the forcing variables $\hat{\sigma}_1$ and $\hat{\sigma}_2$. To estimate the conditional variance $\upsilon(\mathbf{x})$ and density $f(\mathbf{x})$, we use very simple, consistent,

estimators. For pilot bandwidths, we use Scott's rule with $d = 2$ for $j = 1, 2$,

$$\hat{h}_j = \hat{\sigma}_j \cdot n^{-1/6}, \tag{2.22}$$

which is roughly optimal for a normal kernel and multivariate normal data. The optimal bandwidth for a uniform kernel would be slightly higher, however some undersmoothing is desirable given that data are rarely as smooth as a multivariate normal.

Denote the set of treated and untreated units within the uniform kernel as $\mathcal{H}_w \equiv \{i : |X_{1i} - x_1| \leq h_1, |X_{2i} - x_2| \leq h_2, W_i = w\}$ for $w \in \{0, 1\}$. Further, let $N_w \equiv \sum_{i \in \mathcal{H}_w} 1$ denote the number of units in $\mathcal{H}_w$. Following the same approach as Imbens and Kalyanaraman (2008), estimate the density at (x_1, x_2) as

$$\hat{f}(x_1, x_2) = \frac{N_0 + N_1}{N \cdot h_1 \cdot h_2}$$

and the conditional variance at (x_1, x_2) as

$$\hat{v}(x_1, x_2) = \frac{1}{N_0 + N_1} \left(\sum_{i \in \mathcal{H}_0} (Y_i - \frac{1}{N_0} \sum_{j \in \mathcal{H}_0} Y_j)^2 + \sum_{i \in \mathcal{H}_1} (Y_i - \frac{1}{N_1} \sum_{j \in \mathcal{H}_1} Y_j)^2 \right).$$

Both estimators are consistent, although not necessarily efficient.

Step 2: Estimation the second derivatives $m_0^{11}(\mathbf{x})$, $m_1^{11}(\mathbf{x})$, $m_0^{22}(\mathbf{x})$, and $m_1^{22}(\mathbf{x})$:

In the scalar case, Imbens and Kalyanaraman (2008) suggest estimating the second derivatives using local quadratic regression. However, to obtain an optimal pilot bandwidth for this initial estimate requires an estimate a cubic local regression. Yang and Tschernig (1999) confront the same difficulty applied to bandwidth selection rules for multivariate regression problems at interior points.

To simplify, we estimate the control and treatment ($w = 0, 1$) Hessian terms using local quadratic regression with a simple rule-of-thumb pilot bandwidth $h_{j,w} = 4 \cdot \sigma_j^{1/4} n_w^{-1/8}$ and bandwidth matrix $\mathbf{H}_w^{1/2} = \text{diag}[\ h_{1,w} \quad h_{2,w}\ ]$. With a uniform kernel, the local quadratic regression is

$$Y_i = \gamma_0 + \gamma_1(X_1 - x_1) + \gamma_2(X_2 - x_2) + \gamma_3(X_1 - x_1)^2 + \gamma_4(X_2 - x_2)^2 + \gamma_5(X_1 - x_1)(X_2 - x_2) + \nu_i$$

for units with $W_i = 1$ and, separately, $W_i = 0$. Using the coefficients based on units with $W_i = w$, we obtain estimates of the second derivatives as $\hat{m}_w^{11}(\mathbf{x}) = 2 \cdot \hat{\gamma}_3$ and $\hat{m}_w^{22}(\mathbf{x}) = 2 \cdot \hat{\gamma}_4$. At the expense of additional complexity, a more careful strategy for estimating the Hessians terms could be constructed following Yang and Tschernig (1999).

Step 3: Calculate $\hat{h}_{\mathbf{opt}}(\mathbf{x}_k)$ for K evenly spaced points.

Choose K points along the boundary. Plugging in the estimates $\hat{\sigma}_1$, $\hat{\sigma}_2$, $\hat{v}(\mathbf{x}_k)$, $\hat{f}(\mathbf{x}_k)$, $\hat{m}_0^{11}(\mathbf{x}_k)$, $\hat{m}_1^{11}(\mathbf{x}_k)$, $\hat{m}_0^{22}(\mathbf{x}_k)$, and $\hat{m}_1^{22}(\mathbf{x}_k)$ into the optimal bandwidth expression (2.21) yields $\hat{h}_{\text{opt}}(\mathbf{x}_k)$ for each candidate point on the boundary.

Step 4: Select the rule-of-thumb bandwidth $h_{\mathbf{ROT}}$.

Select the minimum $\hat{h}_{\text{opt}}(\mathbf{x}_k)$ as the final rule-of-thumb bandwidth h_{ROT}.

2.5 Application

2.5.1 Data

To demonstrate these methods, we use data from Jacob and Lefgren (2004) on a summer school and grade retention accountability policy instituted in 1996 by the Chicago Public Schools (CPS). This data has also been used by Frandsen and Guiteras (2010) to estimate distributional effects from RD designs. The CPS policy assigned children who failed either a June math or reading exam to a six-week summer school. At the end of the summer, in August, children retook the exams to determine whether

they should be retained or promoted. Students that scored below a cutoff in either reading or math using the best score they obtained across their August or June tests were recommended for retention.

Children in third, sixth, and eighth grade faced the accountability policy. We use data for third graders only, from 1997 to 1999. Jacob and Lefgren (2004) also study the accountability policy's impact on sixth graders and include pre-accountability policy data as an additional control in their analysis. For third graders, the cutoff for both summer school and retention was 2.8 grade equivalents on the reading and math components of the Iowa Test of Basic Skills (ITBS), which corresponds to roughly the 20th percentile in the national achievement distribution. Figure 2.1 illustrates the resulting treatment assignment trajectories. We consider the summer school assignment only. The estimated effects on test scores one year later are therefore the cumulative effect of summer school enrollment and potentially being retained.

After dropping children with missing information, there are 70,831 third graders exposed to the accountability policy. Outcomes one year later are reported as Rasch test scores rather than grade equivalents. Rasch test scores are normalized for the population of third-grade students from 1997 to 1999. To counteract rounding, we add 0.1 of uniform noise to each score when calculating the optimal bandwidth. Jacob and Lefgren (2004) also consider test scores measured two years after summer school and sixth graders. For brevity, we do not report fuller results.

2.5.2 Results

2.5.2.1 First stage

Most students comply with the accountability policy. Figure 2.2 plots June reading and math scores for 2,000 random students that attend summer school (left) or don't (right). While some students do not follow the assignment policy, particularly if assigned to summer school, the treatment assignment rule is quite sharp.

Figure 2.4 performs the same type of analysis on the full sample but uses local linear regression to estimate the probability of attending summer school on a grid of points. The estimated optimal kernel bandwidth, calculated using the rule of thumb

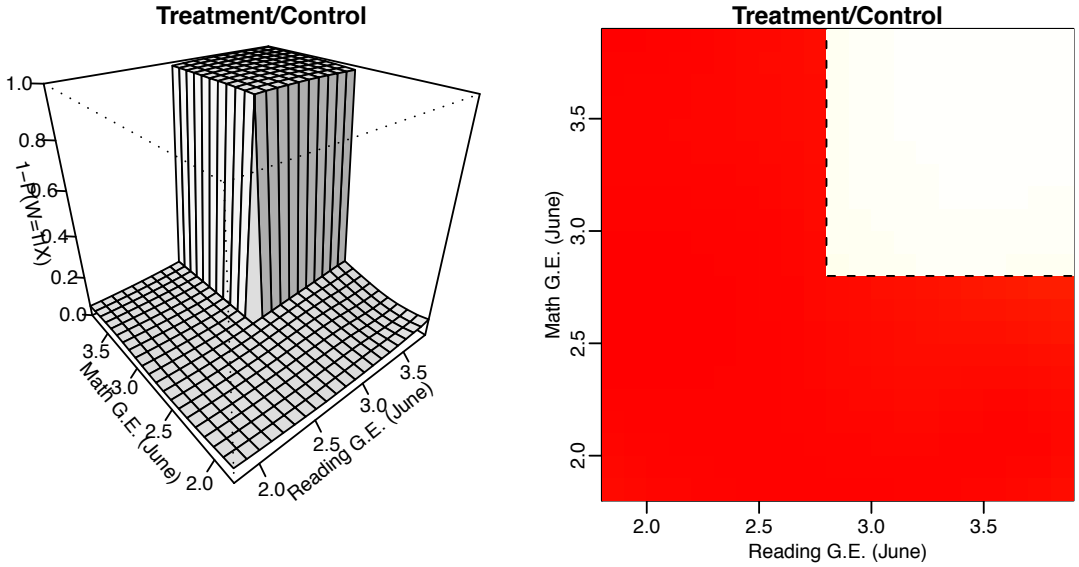


Figure 2.4: Local linear estimates of summer school attendance by June test scores. The local linear estimates use a product edge kernel with bandwidth $h = 0.65$, chosen by the rule of thumb selection rule.

selection rule with $K = 36$, is $h = 0.65$. To visualize higher-dimensional RD designs, we experiment with both a three-dimensional plot (left) and a heatmap (right). Both representations show a sharp jump in the probability of attending summer schools for students that fail one or more subjects.

2.5.2.2 Conditional effects

Figure 2.5 plots local linear estimates of test scores one year later as a function of baseline scores in reading and math. The optimal bandwidth selection rule yields $h = 0.55$ for math and $h = 0.62$ for reading. The small downward dip along the boundary suggests that summer school has a positive effect for students along the boundary.

While three-dimensional plots and heatmaps can help visualize RD outcomes given a two-dimensional boundary, it is difficult to see heterogeneity along the boundary or estimation uncertainty. Figure 2.6 plots the estimated sharp conditional effect, $\tau_{\text{SBRD}}(\mathbf{x})$, on reading and and math scores a year later as a function of baseline

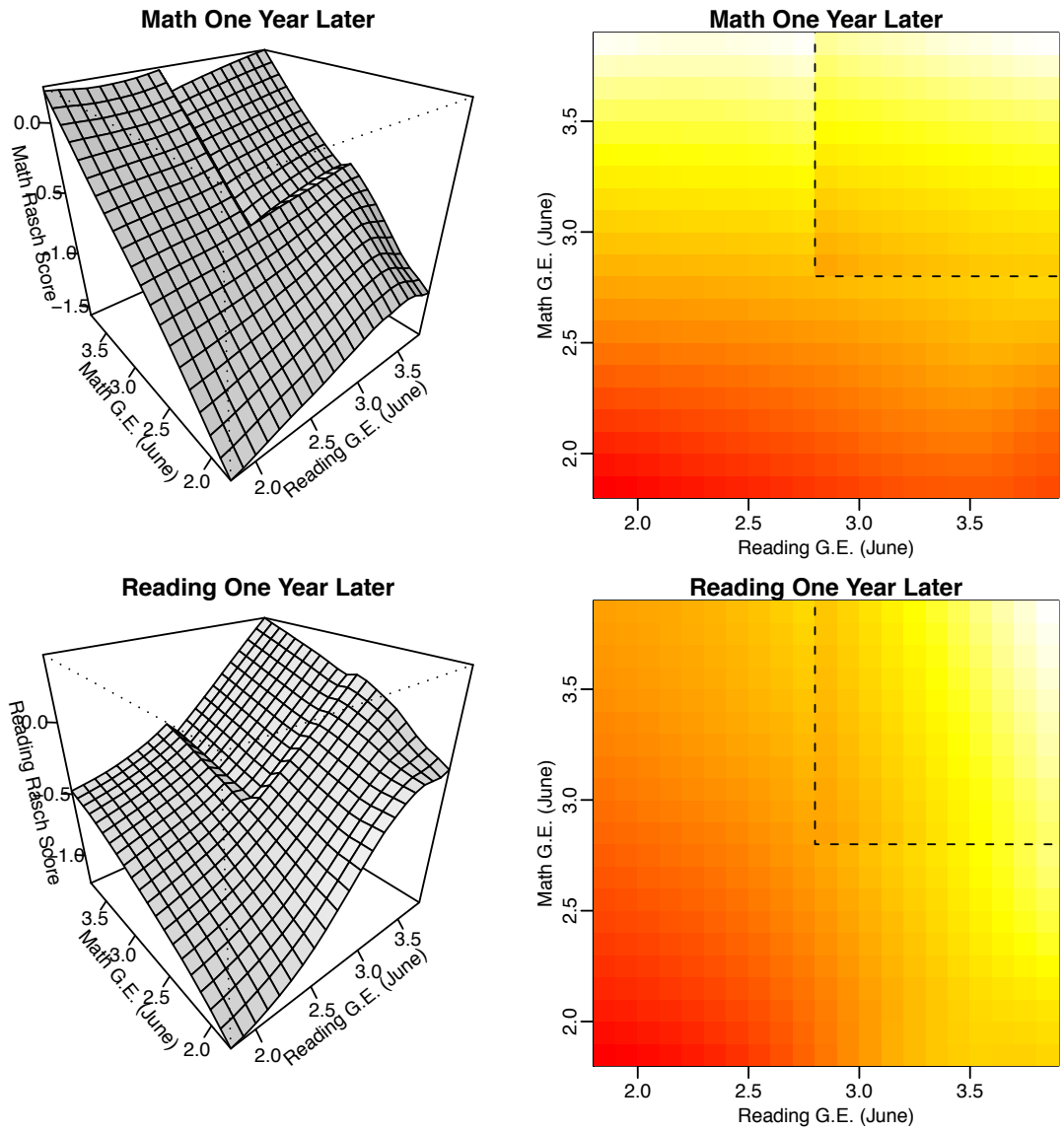


Figure 2.5: Math and reading outcomes one year later by baseline reading and math score. Local linear regression estimates use a product edge kernel with a bandwidth $h = 0.55$ for math and $h = 0.62$ for reading, chosen using the rule-of-thumb selection rule.

scores. The left panel shows the effect of failing reading for students that pass math; the right panel shows the effect of failing math for students that pass reading. While we split the graphs for presentational purposes, all estimates are based on the full data.

Students along both boundary segments score roughly 0.1 to 0.2 standard deviations higher in math a year later. For reading, the results are more mixed. Students that score well in math but just fail reading appear to benefit more in terms of reading outcomes than students that score poorly in math. As seen in the right panel of Figure 2.6, we cannot reject zero effect on reading outcomes for children who just fail math but already score well in reading, although the confidence intervals are wide.

Figure 2.7 plots the fuzzy conditional effects, $\tau_{\text{FBRD}}(\mathbf{x})$, for reading and math scores one year later. The only difference between Figure 2.6 and 2.7 is that Figure 2.7 corrects for different levels of compliance along the boundary whereas Figure 2.6 reports the intent-to-treat effect. Given the high compliance rate, Figures 2.6 and 2.7 are similar. Higher performing math students that just fail reading (bottom, left), seem to gain more in terms reading. Students that just fail math (right), experience statistically significant positive gains provided they don't score too highly on the non-failing reading dimension. The large confidence intervals for high performing reading students (right) is due to the sparsity on this part of the boundary (see Figure 2.2).

2.5.2.3 Average effects

We estimate average effects first using "distance to the nearest boundary" as a scalar forcing variable and select the optimal bandwidth following Imbens and Kalyanaraman (2008). Figure 2.8 gives the standard scalar RD graphs, including both binned means and the local linear fit, for the first-stage treatment probability, and reading and math outcomes one year later. Averaged over the boundary, students just failing either subject are 66 percentage points more likely to attend summer school. Most noncompliance occurs near the boundary. One year later, students that just fail under the accountability policy score 0.14 standard deviations higher in math and 0.07 standard deviations higher in reading. Both effects are statistically significant.

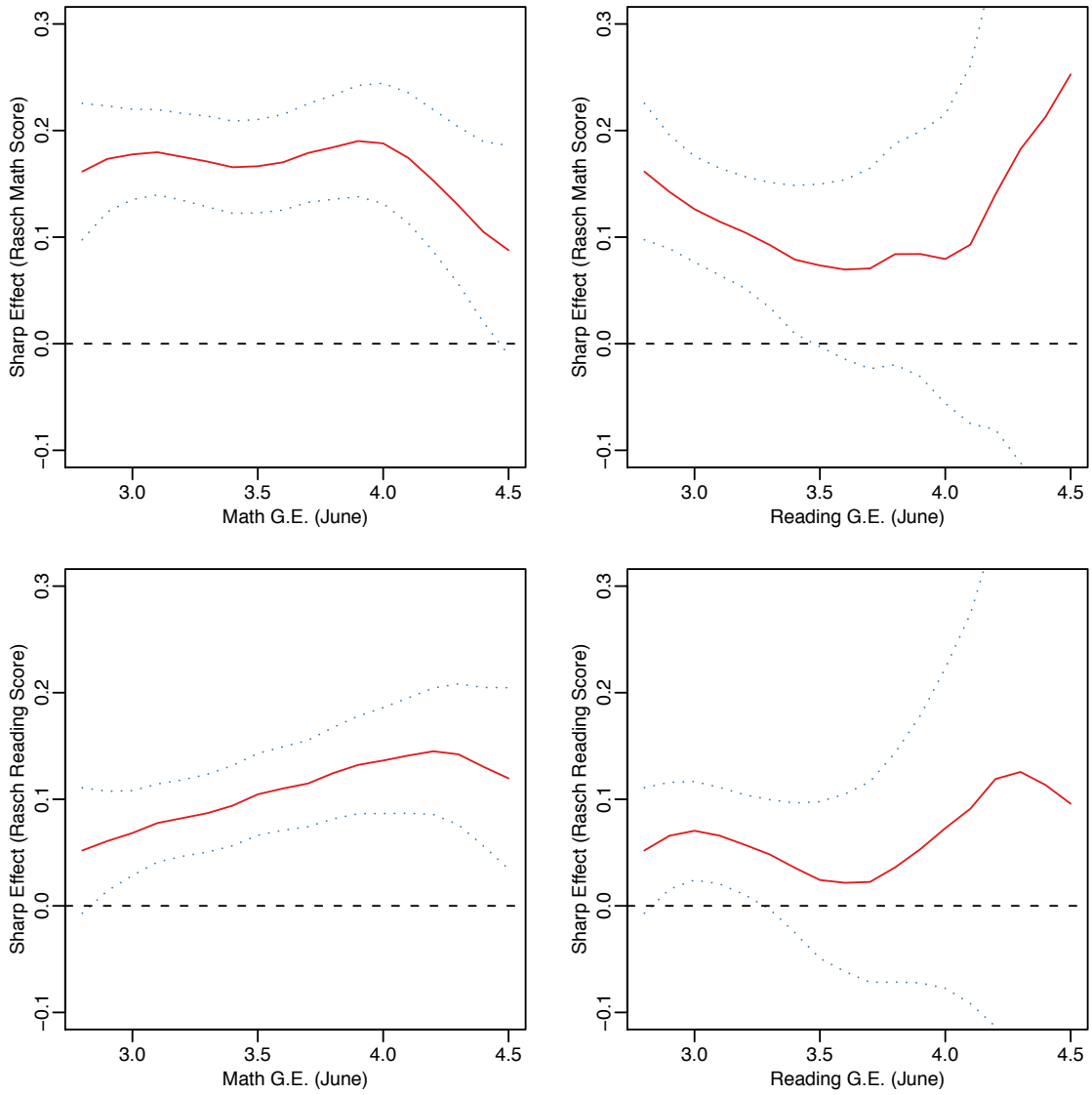


Figure 2.6: Math (top) and reading (bottom) outcomes one year later, $\tau_{\text{SBRD}}(\mathbf{x})$, conditional on baseline math (left) and reading (right) score. The left column represents children on the failing reading boundary; the right column represents children on the failing math boundary. Dotted lines give 95% pointwise confidence intervals.

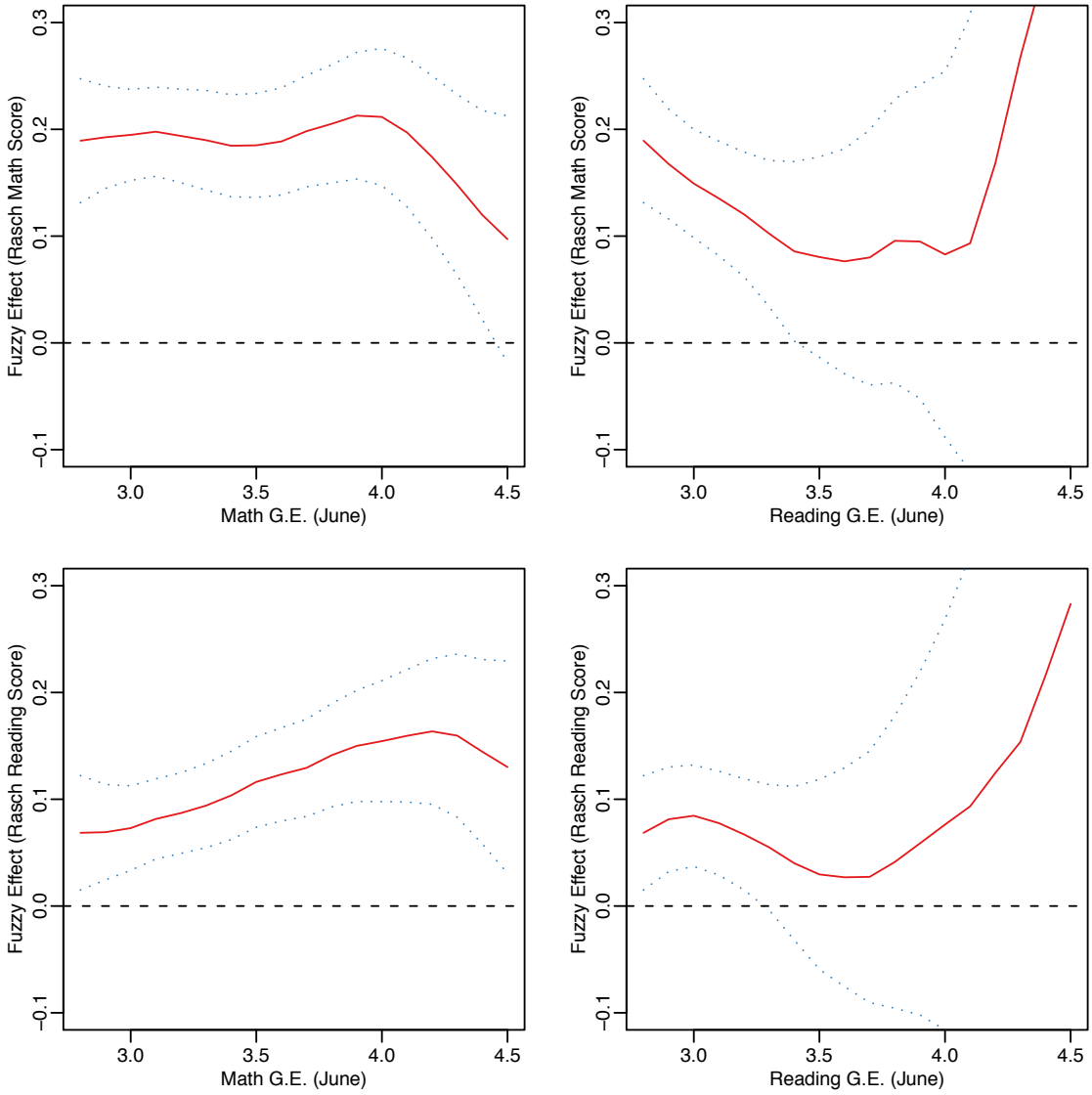


Figure 2.7: Complier math (top) and reading (bottom) outcomes one year later, $\tau_{\text{FBRD}}(\mathbf{x})$, by baseline math (left) and reading (right) score.

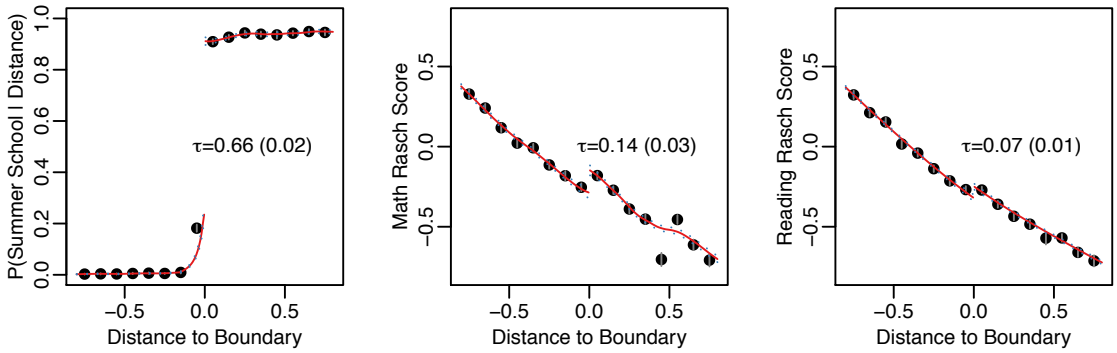


Figure 2.8: Sharp average effects, τ_{SBRD}, for treatment and math and reading scores one year later. Estimates use distance to the nearest boundary as a scalar forcing variable and an edge kernel with bandwidths $h_{\text{treat}} = 0.14$, $h_{\text{math}} = 0.24$, and $h_{\text{read}} = 0.62$, selected following Imbens and Kalyanaraman (2008).

As a sensitivity check, Figure 2.9 graphs the fuzzy average effect for different bandwidth choices. The fuzzy effect estimates the average effect along the boundary for students that comply with the accountability policy. For math, the estimated effect drops somewhat as the bandwidth increases but is consistently between 0.15 and 0.2 standard deviations. For reading, the estimated average fuzzy effect stays constant at 0.09 for bandwidths ranging from 0.2 to 1.5.

As an alternative to using distance to the nearest boundary as a scalar forcing variable, we average the conditional effect over the boundary explicitly, following (2.19), and calculate a standard error using the nonparametric delta method variance estimate given by Theorem 2.A.4. We integrate along the boundary between 2.8 and 5.5 in both directions in intervals of 0.5 ($K = 11$), 0.25 ($K = 21$), and 0.1 ($K = 55$). For scalar RD, we use the optimal bandwidth selected following Imbens and Kalyanaraman (2008); for the averaged conditional effects we use the optimal bandwidth selected following Section 2.4. Table 2.1 compares these different estimates of average effects. All the procedures yield similar results for the sharp average treatment effect. Given the added complexity of averaging explicitly and the importance of presenting RD estimates graphically, we therefore recommend using distance to the boundary to compute average effects.

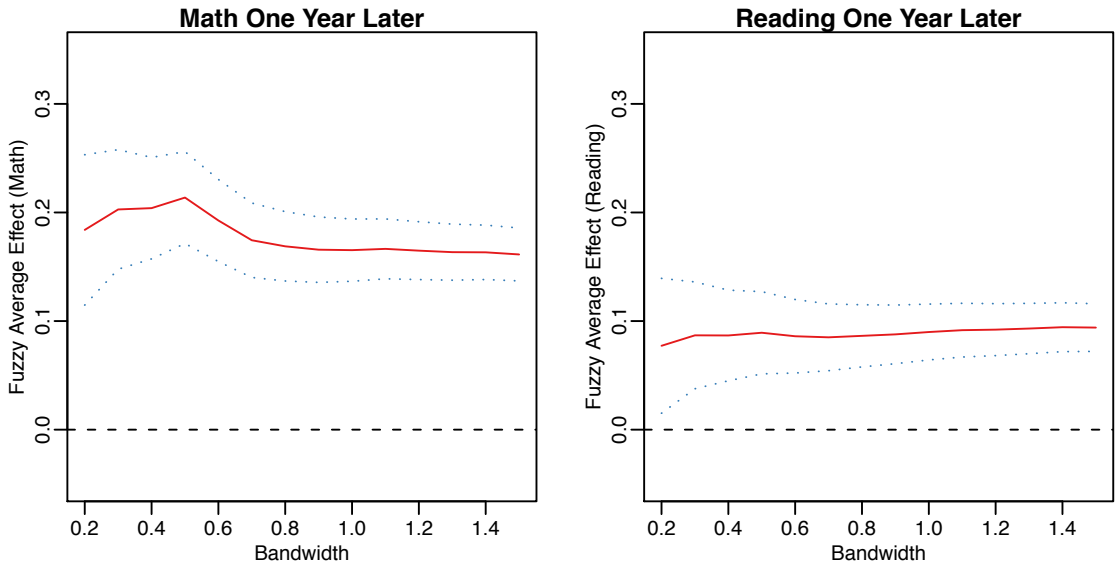


Figure 2.9: Fuzzy average effects, τ_{FBRD}, for math and reading scores one year later using different bandwidths and distance to the nearest boundary as a scalar forcing variable.

		Averaged Conditional Effects		
Sharp Average Treatment Effect	**Scalar RD**	**K = 11**	**K = 21**	**K = 55**
Math one year later	0.14	0.15	0.15	0.15
	(0.025)	(0.015)	(0.014)	(0.014)
Reading one year later	0.07	0.09	0.09	0.09
	(0.014)	(0.015)	(0.014)	(0.014)

Table 2.1: Comparison between averaging conditional effects explicitly and using distance to the nearest boundary as a scalar forcing variable. Explicit averaging uses the nonparametric delta method to calculate standard errors; scalar local linear regression uses robust standard errors. K denotes the number of integration points along the boundary.

2.6 Conclusion

Treatment assignment rules often depend on a vector of forcing variables. Policies with vector forcing variables, or a continuous non-forcing covariate, provide the opportunity to estimate credible causal effects along the entire treatment assignment boundary and explore treatment effect heterogeneity.

We apply simple extensions of ideas developed in the context of scalar RD design to boundary RD designs. We discuss estimation of both conditional sharp and fuzzy effects using multiple local linear regression. As in the scalar case, multiple local linear regression for conditional effects can be framed as weighted OLS or weighted TSLS, making both estimation and inference straightforward. A major issue for all nonparametric techniques is the degree of smoothing. We derive an optimal, data-dependent, bandwidth selection rule for the leading two-dimensional RD case.

We also discuss estimating average effects by integrating the conditional effects along the boundary or by reducing higher-dimensional RD designs to their scalar counterpart. We recommend the latter approach because RD graphs are central to RD design's appeal but become cumbersome in higher dimensions. For the explicitly integrated average effect, we give an estimator for the variance based on the nonparametric delta method.

Our application to Chicago's remedial education policy demonstrates how boundary RD designs can exploit variation along the entire boundary to obtain more detailed knowledge of treatment effect heterogeneity.

2.A Appendix

2.A.1 Proofs

Proof of Theorem 2.2.7.

Proof. Follows directly as in Hahn et al. (2001) but with alternate notation. Consider the numerator and denominator. First,

$$\lim_{\varepsilon \to 0} \mathbb{E}\left[W \mid \mathbf{X} \in N_{\epsilon}^{+}(\mathbf{x})\right] - \lim_{\varepsilon \to 0} \mathbb{E}\left[W \mid \mathbf{X} \in N_{\epsilon}^{-}(\mathbf{x})\right]$$

$$= \lim_{\varepsilon \to 0} \mathbb{E}\left[W \mid \mathbf{X} \in N_{\epsilon}^{+}(\mathbf{x}), Z = 1\right] - \lim_{\varepsilon \to 0} \mathbb{E}\left[W \mid \mathbf{X} \in N_{\epsilon}^{-}(\mathbf{x}), Z = 0\right]$$

$$= \lim_{\varepsilon \to 0} \mathbb{E}\left[W(1) \mid \mathbf{X} \in N_{\epsilon}^{+}(\mathbf{x})\right] - \lim_{\varepsilon \to 0} \mathbb{E}\left[W(0) \mid \mathbf{X} \in N_{\epsilon}^{-}(\mathbf{x})\right]$$

$$= \mathbb{E}\left[W(1) - W(0) \mid \mathbf{X} = \mathbf{x}\right].$$

Second,

$$\lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_{\epsilon}^{+}(\mathbf{x})\right] - \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_{\epsilon}^{-}(\mathbf{x})\right]$$

$$= \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_{\epsilon}^{+}(\mathbf{x}), Z = 1\right] - \lim_{\varepsilon \to 0} \mathbb{E}\left[Y \mid \mathbf{X} \in N_{\epsilon}^{-}(\mathbf{x}), Z = 0\right]$$

$$= \lim_{\varepsilon \to 0} \mathbb{E}\left[Y(0) \cdot (1 - W(1)) + Y(1) \cdot W(1) \mid \mathbf{X} \in N_{\epsilon}^{+}(\mathbf{x})\right]$$

$$- \lim_{\varepsilon \to 0} \mathbb{E}\left[Y(0) \cdot (1 - W(0)) + Y(1) \cdot W(0) \mid \mathbf{X} \in N_{\epsilon}^{-}(\mathbf{x})\right]$$

$$= \mathbb{E}\left[(Y(1) - Y(0)) \cdot (W(1) - W(0)) \mid \mathbf{X} = \mathbf{x}\right].$$

Finally, by the fact that monotonicity rules out $W(1) - W(0) = -1$,

$$\frac{\mathbb{E}\left[(Y(1) - Y(0)) \cdot (W(1) - W(0)) \mid \mathbf{X} = \mathbf{x}\right]}{\mathbb{E}\left[W(1) - W(0) \mid \mathbf{X} = \mathbf{x}\right]}$$

$$= \frac{\mathbb{E}\left[Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}, W(1) - W(0) = 1\right]}{\mathbb{E}\left[W(1) - W(0) \mid \mathbf{X} = \mathbf{x}\right]} \Pr\left(W(1) - W(0) = 1 \mid \mathbf{X} = \mathbf{x}\right)$$

$$= \mathbb{E}\left[Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}, W(1) > W(0)\right] \equiv \tau_{\text{VFRD}}(\mathbf{x}).$$

□

Identification boundary density for compliers $\lambda(\mathbf{x})$.

Proof. By the law of total probability and monotonicity (no defiers),

$$f\left(\mathbf{x} \mid W(1) > W(0), \mathbf{X} \in \mathbb{B}\right) = \frac{1}{\Pr\left(W(1) > W(0) \mid \mathbf{X} \in \mathbb{B}\right)} \cdot [f(\mathbf{x} \mid \mathbf{X} \in \mathbb{B})$$

$$- f(\mathbf{x} \mid W(0) = W(1) = 0, \mathbf{X} \in \mathbb{B}) \cdot \Pr(W(0) = W(1) = 0 \mid \mathbf{X} \in \mathbb{B})$$
$$- f(\mathbf{x} \mid W(0) = W(1) = 1, \mathbf{X} \in \mathbb{B}) \cdot \Pr(W(0) = W(1) = 1, \mathbf{X} \in \mathbb{B})].$$

Following the same logic as, for instance, Abadie (2003) Lemma 3.1, the proportion of compliers, always-takers, and never-takers along the boundary are identified as

$$\Pr\left(W(1) > W(0) \mid \mathbf{X} \in \mathbb{B}\right) = \lim_{\epsilon \to 0} \mathbb{E}[W \mid \mathbf{X} \in B_\epsilon^+] - \lim_{\epsilon \to 0} \mathbb{E}[W \mid \mathbf{X} \in B_\epsilon^-],$$

$$\Pr\left(W(1) = W(0) = 0 \mid \mathbf{X} \in \mathbb{B}\right) = 1 - \lim_{\epsilon \to 0} \mathbb{E}[W \mid \mathbf{X} \in B_\epsilon^+]$$

$$\Pr\left(W(1) = W(0) = 1 \mid \mathbf{X} \in \mathbb{B}\right) = \lim_{\epsilon \to 0} \mathbb{E}[W \mid \mathbf{X} \in B_\epsilon^-]$$

The conditional densities can be identified similarly. However, both densities are at boundary points. Let $\mathbf{x}_1$ be a sequence that converges to $\mathbf{x}$ from within N_ϵ^+ and $\mathbf{x}_0$ be a sequence that converges to $\mathbf{x}$ from within N_ϵ^-. Then, for all $\mathbf{x} \in \mathbb{B}$,

$$\begin{aligned} f(\mathbf{x} \mid W(0) = W(1) = 0) &= \lim_{\mathbf{x}_1 \to \mathbf{x}} f(\mathbf{x}_1 \mid W(0) = W(1) = 0) \\ &= \lim_{\mathbf{x}_1 \to \mathbf{x}} f(\mathbf{x}_1 \mid W = 0), \\ f(\mathbf{x} \mid W(0) = W(1) = 1) &= \lim_{\mathbf{x}_0 \to \mathbf{x}} f(\mathbf{x}_0 \mid W(0) = W(1) = 1) \\ &= \lim_{\mathbf{x}_0 \to \mathbf{x}} f(\mathbf{x}_0 \mid W = 1). \end{aligned}$$

The conditional densities follow from

$$f\left(\mathbf{x} \mid \mathbf{X} \in \mathbb{B}\right) = \frac{f\left(\mathbf{x}\right)}{\int_{\mathbf{x} \in \mathbb{B}} f\left(\mathbf{x}\right) d\mathbf{x}}.$$

Specifically,

$$\begin{aligned} f\left(\mathbf{x} \mid W(1) > W(0), \mathbf{X} \in \mathbb{B}\right) = \frac{1}{\lim_{\epsilon \to 0} \mathbb{E}[W \mid \mathbf{X} \in B_\epsilon^+] - \lim_{\epsilon \to 0} \mathbb{E}[W \mid \mathbf{X} \in B_\epsilon^-]} \cdot [\frac{f\left(\mathbf{x}\right)}{\int_{\mathbf{x} \in \mathbb{B}} f\left(\mathbf{x}\right) d\mathbf{x}} \\ - \frac{\lim_{\mathbf{x}_1 \to \mathbf{x}} f(\mathbf{x}_1 \mid W = 0)}{\int_{\mathbf{x} \in \mathbb{B}} \lim_{\mathbf{x}_1 \to \mathbf{x}} f(\mathbf{x}_1 \mid W = 0) d\mathbf{x}} \cdot \left(1 - \lim_{\epsilon \to 0} \mathbb{E}[W \mid \mathbf{X} \in B_\epsilon^+]\right) \\ - \frac{\lim_{\mathbf{x}_0 \to \mathbf{x}} f(\mathbf{x}_0 \mid W = 1)}{\int_{\mathbf{x} \in \mathbb{B}} \lim_{\mathbf{x}_0 \to \mathbf{x}} f(\mathbf{x}_0 \mid W = 1) d\mathbf{x}} \cdot \lim_{\epsilon \to 0} \mathbb{E}[W \mid \mathbf{X} \in B_\epsilon^-]] \equiv \lambda(\mathbf{x}). \quad (2.23) \end{aligned}$$

In practice, $\lambda(\mathbf{x})$ can be estimated by a combination of local averages and kernel density estimates with potential boundary corrections (e.g., Jones, 1993; Hjort and Jones, 1996; Loader, 1996). □

2.A.2 Nonparametric delta method variance estimator

Consider a statistical functional $\theta(\mathbf{x}) = T(\mathbf{x}; F)$ where T is a function of the cumulative distribution function $F(\mathbf{z})$ and can be indexed by a parameter $\mathbf{x}$. For instance $T(\mathbf{x}; F)$ could be a mean, $T(\mathbf{x}; F) = \int z_1 dF(\mathbf{z})$, or a kernel density estimate at $\mathbf{x}$, $T(\mathbf{x}; F) = \int K_{\mathbf{H}}(\mathbf{z}-\mathbf{x})dF(\mathbf{z})$. A plug-in estimator for θ is $\hat{\theta}_n(\mathbf{x}) = T(\mathbf{x}; \hat{F}_n)$ where $\hat{F}_n$ is the empirical CDF of $\mathbf{z}$. In what follows we sometimes suppress $\mathbf{x}$ for notational convenience. However, it is helpful to explicitly indicate the parameter $\mathbf{x}$ for kernel density and local linear regression estimates evaluated at a specific point.

The *influence function* for $T(F)$ is defined as

$$L_T(\mathbf{z}; F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\delta_{\mathbf{z}}) + T(F)}{\epsilon} = \left. \frac{\partial T((1-\epsilon)F + \epsilon\delta_{\mathbf{z}})}{\partial \epsilon} \right|_{\epsilon=0},$$

where $\delta_{\mathbf{z}}$ is a point mass at $\mathbf{z}$. The *empirical influence function* is $L_T(\mathbf{Z}_i; \hat{F}_n)$. The *nonparametric delta method variance estimate* for θ is then

$$\hat{\nu}^2 = \frac{1}{n} \sum_{i=1}^{n} L_T(\mathbf{Z}_i; \hat{F}_n)^2$$

Asymptotically valid $1-\alpha$ confidence intervals are given by $T(\hat{F}_n) \pm z_{\alpha/2} \cdot \hat{\nu}/\sqrt{n}$. For reviews of this influence function approach to variance estimation see, for instance, Davison and Hinkley (1997) or Wasserman (2006).

The sharp average RD parameter can be expressed explicitly in terms of kernel estimators

$$\begin{aligned}\tau_{\text{SBRD}}(F(Y,\mathbf{X},W)) &= \int_{\mathbf{x}\in\mathbb{B}} \tau_{\text{SBRD}}(\mathbf{x};F)\cdot f(\mathbf{x};F \mid \mathbf{x}\in\mathbb{B})d\mathbf{x} \\ &= \left(\int_{\mathbf{x}\in\mathbb{B}} f(\mathbf{x};F)d\mathbf{x}\right)^{-1}\int_{\mathbf{x}\in\mathbb{B}} \tau_{\text{SBRD}}(\mathbf{x};F)\cdot f(\mathbf{x};F)d\mathbf{x},\end{aligned}$$

where $\tau_{\text{VSRD}}(\mathbf{x};F)$ and $f(\mathbf{x};F)$ are the population quantities for the local linear regression and kernel density at $\mathbf{x}$ over $F(Y,\mathbf{X},W)$. The influence function for the average effect $\tau_{\text{SBRD}}(F(Y,\mathbf{X},W))$ depends on the influence functions for the conditional effect $\tau_{\text{SBRD}}(\mathbf{x};F)$ and density $f(\mathbf{x};F)$. We therefore consider these first:

Lemma 2.A.1. (Influence function of a Kernel Density Estimate) *The empirical influence function for a kernel density estimate at* $\mathbf{x}$*,* $f(\mathbf{x};F) = \mathbb{E}\left[K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x})\right]$*, is*

$$L_{f(\mathbf{x})}(\mathbf{X}_i;\hat{F}) = K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x}) = K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x}) - \hat{f}(\mathbf{x}).$$

Proof. Follows immediately from definition or from that kernel densities are linear statistics (e.g., Davison and Hinkley, 1997, p. 47). □

Lemma 2.A.2. (Influence Function of Conditional Sharp Treatment Effect) *Let* $\mathbf{R}_i$ *be the vector of covariates defined in* (2.13) *and* $\mathbf{B}$ *be the* $n\times n$ *diagonal weight matrix with diagonal elements* $b_i = K_{\mathbf{H}}(\mathbf{X}_i-\mathbf{x})$*, both functions of* $\mathbf{x}$*. Then the empirical influence function for the conditional sharp treatment effect at* $\mathbf{x}$ *is*

$$L_{\tau_{SBRD}(\mathbf{x})}(Y_i,\mathbf{R}_i,b_i;\hat{F}) = n\cdot\mathbf{e}_2'\left(\mathbf{R}'\mathbf{B}\mathbf{R}\right)^{-1}\mathbf{R}_i b_i\left(Y_i-\mathbf{R}_i'\hat{\boldsymbol{\beta}}\right),$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{R}'\mathbf{B}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{B}\mathbf{Y}$.

Proof. From the definition of a local linear estimation and influence function,

$$L_{\tau_{\text{SBRD}}(\mathbf{x})}(Y_i, \mathbf{R}_i, b_i; \hat{F}) = \left(\frac{\partial}{\partial \epsilon} \mathbf{e}_2' \left[n^{-1}(1-\epsilon) \cdot \mathbf{R}'\mathbf{B}\mathbf{R} + \epsilon \cdot \mathbf{R} b_i \mathbf{R}_i' \right]^{-1} \left[n^{-1}(1-\epsilon) \cdot \mathbf{R}'\mathbf{B}\mathbf{Y} + \epsilon \cdot \mathbf{R}_i' b_i Y_i \right] \right) \Bigg|_{\epsilon=0}$$

$$= \mathbf{e}_2' \left(-n \left[\mathbf{R}'\mathbf{B}\mathbf{R} \right]^{-1} \left[-n^{-1}\mathbf{R}'\mathbf{B}\mathbf{R} + \mathbf{R}_i b_i \mathbf{R}_i' \right] n \left[\mathbf{R}'\mathbf{B}\mathbf{R} \right]^{-1} n^{-1}\mathbf{R}'\mathbf{B}\mathbf{Y} + n \left[\mathbf{R}'\mathbf{B}\mathbf{R} \right]^{-1} \left[-n^{-1}\mathbf{R}'\mathbf{B}\mathbf{Y} + \mathbf{R}_i' b_i Y_i \right] \right)$$

$$= \mathbf{e}_2' \left(\left[\mathbf{I} - n \left[\mathbf{R}'\mathbf{B}\mathbf{R} \right]^{-1} \mathbf{R}_i b_i \mathbf{R}_i' \right] \hat{\boldsymbol{\beta}} + \left[-\hat{\boldsymbol{\beta}} + n \left[\mathbf{R}'\mathbf{B}\mathbf{R} \right]^{-1} \mathbf{R}_i' b_i Y_i \right] \right)$$

$$= n \cdot \mathbf{e}_2' \left[\mathbf{R}'\mathbf{B}\mathbf{R} \right]^{-1} \mathbf{R}_i b_i \left(Y_i - \mathbf{R}_i' \hat{\boldsymbol{\beta}} \right).$$

□

Theorem 2.A.3. (Influence Function of Sharp average Treatment Effect) *Let* $\mathbf{z} = (y, \mathbf{r}, b)$ *and* $\pi = \int_{\mathbf{x} \in \mathbb{B}} f(\mathbf{x}; F) d\mathbf{x}$. *The influence function for the sharp average treatment effect* τ_{VSRD} *is*

$$L_{\tau_{SBRD}}(\mathbf{z}; F) = \frac{1}{\pi} \int_{\mathbf{x} \in \mathbb{B}} \left(L_{\tau_{SBRD(\mathbf{x})}}(\mathbf{z}; F) \cdot f(\mathbf{x}; F) + \left[\tau_{SBRD}(\mathbf{x}; F) - \tau_{SBRD}(F) \right] \cdot L_{f(\mathbf{x})}(\mathbf{z}; F) \right) d\mathbf{x}.$$

Proof. Follows from the influence function definition:

$$L_{\tau_{\text{SBRD}}}(\mathbf{z}; F) = \frac{\partial}{\partial \epsilon} \frac{\int_{\mathbf{x}\in\mathbb{B}} \tau_{\text{SBRD}}(\mathbf{x}; (1-\epsilon)\cdot F + \epsilon \cdot \delta_z) \cdot f(\mathbf{x}; (1-\epsilon)\cdot F + \epsilon \cdot \delta_z) d\mathbf{x}}{\int_{\mathbf{x}\in\mathbb{B}} f(\mathbf{x}; (1-\epsilon)\cdot F + \epsilon \cdot \delta_z) d\mathbf{x}} \Bigg|_{\epsilon=0}$$

$$= -\frac{1}{\pi^2} \cdot \int_{\mathbf{x}\in\mathbb{B}} L_{f(\mathbf{x})}(\mathbf{z}; F) d\mathbf{x} \cdot \int_{\mathbf{x}\in\mathbb{B}} \tau_{\text{SBRD}}(\mathbf{x}; F) \cdot f(\mathbf{x}; F) d\mathbf{x}$$

$$+ \frac{1}{\pi} \int_{\mathbf{x}\in\mathbb{B}} \left(L_{\tau_{\text{SBRD}}(\mathbf{x})}(\mathbf{z}; F) \cdot f(\mathbf{x}; F) + \tau_{\text{SBRD}}(\mathbf{x}; F) \cdot L_{f(\mathbf{x})}(\mathbf{z}; F) \right) d\mathbf{x}$$

$$= \frac{1}{\pi} \int_{\mathbf{x}\in\mathbb{B}} \left(L_{\tau_{\text{SBRD}}(\mathbf{x})}(\mathbf{z}; F) \cdot f(\mathbf{x}; F) + \tau_{\text{SBRD}}(\mathbf{x}; F) \cdot L_{f(\mathbf{x})}(\mathbf{z}; F) - \tau_{\text{SBRD}}(F) \cdot L_{f(\mathbf{x})}(\mathbf{z}; F) \right) d\mathbf{x}$$

$$= \frac{1}{\pi} \int_{\mathbf{x}\in\mathbb{B}} \left(L_{\tau_{\text{SBRD}}(\mathbf{x})}(\mathbf{z}; F) \cdot f(\mathbf{x}; F) + [\tau_{\text{SBRD}}(\mathbf{x}; F) - \tau_{\text{SBRD}}(F)] \cdot L_{f(\mathbf{x})}(\mathbf{z}; F) \right) d\mathbf{x}.$$

Here $L_{\tau_{\text{SBRD}}(\mathbf{x})}(\mathbf{z}; F)$ and $L_{f(\mathbf{x})}(\mathbf{z}; F)$ are the influence functions the conditional effect and density. □

The influence function of the average effect, and therefore the variance, does not depend on the influence function of the density if there is no treatment heterogeneity and $\tau_{\text{SBRD}}(\mathbf{x}; F) = \tau_{\text{SBRD}}(F)$. Intuitively, it doesn't matter how we weight the conditional effects if they are constant.

Theorem 2.A.4. (Variance Estimate of Sharp average Treatment Effect) *Let $k = 1, \ldots, K$ index evenly spaced points $\mathbf{x}_k$ on the boundary and k subscripts denote definitions relative to $\mathbf{x}_k$, e.g., $b_{k,i} = K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}_k)$. Then, as K becomes large, an estimate for the empirical influence function of the sharp average effect $\hat{\tau}_{VSRD}$ is*

$$\psi_i = \frac{1}{\hat{\pi}} \sum_{k=1}^{k} \left[n \cdot \mathbf{e}_2' \left(\mathbf{R}_k' \mathbf{B}_k \mathbf{R}_k \right)^{-1} \mathbf{R}_{k,i} b_{k,i} \left(Y_i - \mathbf{R}_{k,i} \hat{\boldsymbol{\beta}}_k \right) \cdot \hat{f}(\mathbf{x}_k) + (\hat{\tau}_{SBRD}(\mathbf{x}_k) - \hat{\tau}_{SBRD}) \cdot \left(b_{k,i} - \hat{f}(\mathbf{x}_k) \right) \right]$$

where

$$\hat{\pi} = \sum_{k=1}^{K} \hat{f}(\mathbf{x}_k),$$

$$\hat{f}(\mathbf{x}_k) = n^{-1} \sum_{i=1}^{n} b_{k,i},$$

$$\begin{aligned} \hat{\boldsymbol{\beta}}_k &= \left(\mathbf{R}_k' \mathbf{B}_k \mathbf{R}_k\right)^{-1} \mathbf{R}_k' \mathbf{B}_k \mathbf{Y}, \\ \hat{\tau}_{SBRD}(\mathbf{x}_k) &= \mathbf{e}_2' \hat{\boldsymbol{\beta}}_k, \end{aligned}$$

$$\hat{\tau}_{SBRD} = \hat{\pi}^{-1} \sum_{k=1}^{K} \hat{\tau}_{SBRD}(\mathbf{x}_k) \cdot \hat{f}(\mathbf{x}_k).$$

An estimate of the variance is

$$\hat{\mathbb{V}}(\hat{\tau}_{SBRD} \mid \mathbf{R}) = \frac{1}{n} \sum_{i=1}^{n} \psi_i^2.$$

2.A.3 Optimal bandwidth selection

Ruppert and Wand (1994) derive the properties of multivariate local linear regression. Here we apply their results to the vector-valued RD problem. Consider the nonparametric regression

$$Y_i = m(\mathbf{X}_i) + \upsilon^{1/2}(\mathbf{X}_i)\varepsilon_i$$

where $\upsilon(\mathbf{x}) = \mathbb{V}(Y_i \mid \mathbf{X}_i = \mathbf{x})$ and ε_i are i.i.d. random variables with mean zero and unit variance and are independent of $\mathbf{X}_i$. The sharp conditional effect is a simple functional of two such regressions at a boundary point.

To analyze the asymptotic properties of multivariate locally weighted least squares, Ruppert and Wand (1994, p. 1349) give regularity conditions that are satisfied by spherically symmetric compact kernels or product kernels based on symmetric univariate kernels. We use an two-dimensional edge kernel

$$K(u_1, u_2) = (1 - |u_1|) \cdot (1 - |u_2|) \cdot \mathbf{1}\{|u_2| \leq 1, |u_2| \leq 1\}$$

with $\mathbf{H}_0 = \mathbf{H}_1 = \text{diag}([\ \sigma_1 h \quad \sigma_2 h\ ])$. We also make the RD assumption 2.2.1, which replaces Ruppert and Wand's Assumption A4. Finally, to avoid the degeneracy that arises if the bias for m_1 and m_0 cancel, we assume

Assumption 2.A.5. *The boundary $\mathbb{B}$ consist of vertical and horizontal segments and*

$$\left(C_1 \cdot \sigma_1^2 \cdot \left[m_0^{11}(\mathbf{x}) - m_1^{11}(\mathbf{x})\right] + C_2 \cdot \sigma_2^2 \cdot \left[m_1^{22}(\mathbf{x}) - m_0^{22}(\mathbf{x})\right]\right) \neq 0$$

for horizontal segments when $C_1 = 5$ and $C_2 = 3$ or for vertical segments when $C_1 = 3$ and $C_2 = 5$.

Now consider a boundary point $\mathbf{x} \in \mathbb{B}$. Let $\mathcal{D}^1_{\mathbf{x},\mathbf{H}_1} = \{\mathbf{z} : (\mathbf{x} + \mathbf{H}_1^{1/2}\mathbf{z}) \in \mathbb{T}\} \cap \text{supp}(K)$ and, likewise, $\mathcal{D}^0_{\mathbf{x},\mathbf{H}_0} = \{\mathbf{z} : (\mathbf{x} + \mathbf{H}_0^{1/2}\mathbf{z}) \in \mathbb{T}^c\} \cap \text{supp}(K)$. That is, $\mathcal{D}^1_{\mathbf{x},\mathbf{H}_1}$ and $\mathcal{D}^0_{\mathbf{x},\mathbf{H}_0}$give the treatment and control points within a bandwidth from $\mathbf{x}$ and within the support of the kernel K. Further, let $\mathbf{v} = [\ 1 \quad \mathbf{u}'\ ]'$ where $\mathbf{u} = [\ u_1 \quad \cdots \quad u_d\ ]'$. Then:

Theorem 2.A.6. *Under regularity conditions (Ruppert and Wand, 1994, p. 1349) and Assumptions 2.2.1 the conditional asymptotic MSE for $\tau_{SBRD}(\mathbf{x})$ is*

$$AMSE(h; \mathbf{x}) = \frac{h^4}{3600} \cdot \left(C_1 \cdot \sigma_1^2 \cdot \left[m_0^{11}(\mathbf{x}) - m_1^{11}(\mathbf{x})\right] + C_2 \cdot \sigma_2^2 \cdot \left[m_1^{22}(\mathbf{x}) - m_0^{22}(\mathbf{x})\right]\right)^2 + \frac{32 \cdot \upsilon(\mathbf{x})}{5 \cdot f(\mathbf{x}) \cdot h^2 \cdot \sigma_1 \cdot \sigma_2 \cdot n} + o_p\{\frac{1}{nh^2} + h^4\}. \quad (2.24)$$

with $C_1 = 5$ and $C_2 = 3$ for horizontal boundaries and $C_1 = 3$ and $C_2 = 5$ for vertical boundaries.

Proof. By Ruppert and Wand (1994) Theorem 2.2 the conditional bias for each limit is

$$\mathbb{E}\left[\hat{m}_j(\mathbf{x}) - m_j(\mathbf{x}) \mid \mathbf{X}\right] = \frac{\mathbf{e}_1' N_{j,\mathbf{x}}^{-1}}{2} \int_{\mathcal{D}^j_{\mathbf{x},\mathbf{H}_j}} \mathbf{v} K(\mathbf{u}) \mathbf{u}' \mathbf{H}_j^{1/2} \mathbf{M}_j(\mathbf{x}) \mathbf{H}_j^{1/2} \mathbf{u} d\mathbf{u} + o_p(\mathrm{tr}(\mathbf{H}_j)),$$

and the conditional variance is

$$\mathbb{V}\left[\hat{m}_j(\mathbf{x}) \mid \mathbf{X}\right] = \left(n^{-1} \left|\mathbf{H}\right|^{-1/2} \mathbf{e}_1' N_{j,\mathbf{x}}^{-1} T_{j,\mathbf{x}} N_{j,\mathbf{x}}^{-1} \mathbf{e}_1 / f(\mathbf{x})\right) \upsilon(\mathbf{x}) \left(1 + o_p(1)\right),$$

where

$$N_{j,\mathbf{x}} = \int_{\mathcal{D}^j_{\mathbf{x},\mathbf{H}_j}} \mathbf{v}' K(\mathbf{u}) d\mathbf{u},$$

$$T_{j,\mathbf{x}} = \int_{\mathcal{D}^j_{\mathbf{x},\mathbf{H}_j}} \mathbf{v}' K^2(\mathbf{u}) d\mathbf{u}.$$

Under our specialized assumptions the squared bias for the conditional sharp effect is

$$\mathbb{E}\left[\hat{\tau}_{\mathrm{SBRD}}(\mathbf{x}) - \tau_{\mathrm{SBRD}}(\mathbf{x}) \mid \mathbf{X}\right]^2 =$$

$$\frac{h^4}{3600} \cdot \left(C_1 \cdot \sigma_1^2 \cdot \left[m_0^{11}(\mathbf{x}) - m_1^{11}(\mathbf{x})\right] + C_2 \cdot \sigma_2^2 \cdot \left[m_1^{22}(\mathbf{x}) - m_0^{22}(\mathbf{x})\right]\right)^2, \quad (2.25)$$

and the conditional variance is

$$\mathbb{V}\left[\hat{\tau}_{\mathrm{SBRD}}(\mathbf{x}) \mid \mathbf{X}\right] = \frac{32 \cdot \upsilon(\mathbf{x})}{5 \cdot f(\mathbf{x}) \cdot h^2 \cdot n \cdot \sigma_1 \cdot \sigma_2}. \quad (2.26)$$

Combining squared bias (2.25) and variance (2.26) yields the conditional asymptotic MSE (2.24). □

Chapter 3

Do Value-Added Estimates Add Value? Accounting for Learning Dynamics[1]

Evaluations of educational programs commonly assume that children's learning persists over time. We illustrate the central role of persistence in estimating and interpreting value-added models of learning. Using data from Pakistani public and private schools, we apply dynamic panel methods that address three key empirical challenges to widely used value-added models: imperfect persistence, unobserved student heterogeneity, and measurement error. Our estimates suggest that only a fifth to a half of learning persists between grades and that private schools increase average achievement by 0.25 standard deviations each year. In contrast, value-added models that assume perfect persistence yield severely downwardly biased and occasionally wrong-signed estimates of the private school effect. Models that ignore unobserved heterogeneity or measurement error produce biased estimates of persistence. These

[1]This chapter is coauthored with Tahir Andrabi, Jishnu Das, and Asim Khwaja. We are grateful to Alberto Abadie, Chris Avery, David Deming, Pascaline Dupas, Brian Jacob, Dale Jorgenson, Elizabeth King, Karthik Muralidharan, David McKenzie, Rohini Pande, Lant Pritchett, Jesse Rothstein, Douglas Staiger, Tara Vishwanath, an anonymous referee, and seminar participants at Harvard, NEUDC and BREAD for helpful comments on drafts of this paper. This research was funded by grants from the Poverty and Social Impact Analysis and Knowledge for Change Program Trust Funds and the South Asia region of the World Bank. The findings, interpretations, and conclusions expressed here are those of the authors and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments they represent.

results have implications for program evaluation and value-added acountability system design.

3.1 Introduction

Models of learning often assume that a child's achievement persists between grades—what a child learns today largely stays with her tomorrow. Yet recent research suggests that treatment effects measured by test scores fade rapidly, both in randomized interventions and observational studies. Kane and Staiger (2008), Jacob et al. (2010), and Rothstein (2010) find that teacher effects dissipate by between 50 and 80 percent over one year. The same pattern holds in several studies of supplemental education programs in developed and developing countries. Currie and Thomas (1995) document the rapid fade out of Head Start's impact in the United States, and Glewwe et al. (2003) and Banerjee et al. (2007) report on education experiments in Kenya and India where over 70 percent of the one-year treatment effect is lost after an additional year. Low persistence may in fact be the norm rather than the exception. It appears to be a central feature of learning.

Low persistence has critical implications for commonly used program evaluation strategies that rest heavily on assumptions about or estimation of persistence. Using primary data on public and private schools in Pakistan, this paper addresses the challenges to value-added evaluation strategies posed by (1) imperfect persistence of achievement, (2) heterogeneity in learning, and (3) measurement error in test scores. We find that ignoring any of these learning dynamics biases estimates of persistence and can dramatically affect estimates of the value-added of private schools.

To fix concepts, consider a simple model of learning, $y^*_{it} = \alpha T_{it} + \beta y^*_{i,t-1} + \eta_i + \upsilon_{it}$, where y^*_{it} is child true (unobserved) achievement in period t, T_{it} is the treatment or program effect in period t, and η_i is unobserved student ability that speeds learning each period. We refer to β, the parameter that links achievement across periods, as persistence. The canonical restricted value-added or gain-score model assumes that $\beta = 1$ (for examples, see Hanushek (2003)). When $\beta < 1$, achievement exhibits conditional mean reversion. Estimates of the treatment or program effect, α, that

assume $\beta = 1$ will be biased if the baseline achievement of the treatment and control groups differs and persistence is imperfect. This has led many researchers to advocate leaving lagged achievement on the right-hand side. However, doing so is not entirely straightforward: if estimated by OLS, omitted heterogeneity that speeds learning, η_i, will generally bias β upward and any measurement error in test scores $y_{i,t-1}$ that proxy true achievement $y^*_{i,t-1}$ will bias β downward. Both the estimate of persistence β and the treatment effect α may remain biased when estimated by standard methods.

To address these concerns, we use three years of data on a panel of children to jointly estimate β and the treatment effect α using techniques from the dynamic panel literature (Arellano and Honore, 2001; Arellano, 2003). There are several findings. First, we find that learning persistence is low: only a fifth to a half of achievement persists between grades. That is, β is between 0.2 and 0.5 rather than closer to 1. These estimates are remarkably similar to those obtained in the United States (Kane and Staiger, 2008; Jacob et al., 2010; Rothstein, 2010). The low persistence we find implies that long-run extrapolations from short-run impacts are fraught with danger. In the model above, the long-run impact of continued treatment is $\alpha/(1-\beta)$; with estimates of β around 0.2 to 0.5, these gains may be much smaller than those obtained by assuming that β is close to 1.[2]

Second, OLS estimates of β are contaminated both by measurement error in test scores and unobserved student-level heterogeneity in learning. Ignoring both biases leads to higher persistence estimates between 0.5 and 0.6; correcting only for measurement error results in estimates between 0.7 and 0.8. In our data, the upward bias on persistence from omitted heterogeneity outweighs measurement error attenuation.

Third, the private schooling effect is highly sensitive to the persistence parameter. Since private schooling is a school input that is continually applied and leads to a large baseline gap in achievement, this is expected. We find that incorrectly assuming

[2]For example, Krueger and Whitmore (2001), Angrist et al. (2002), Krueger (2003), and Gordon et al. (2006) calculate the economic return of various educational interventions by citing research linking test scores to earnings of young adults (e.g. Murnane et al., 1995; Neal and Johnson, 1996). Although effects on learning as measured by test-scores may fade, non-cognitive skills that are rewarded in the labor market could persist. For instance, Currie and Thomas (1995), Schweinhart et al. (2005), and Deming (2009) provide evidence of long run effects of Head Start and the Perry Preschool Project, even though cognitive gains largely fade after children enroll in regular classes.

$\beta = 1$ significantly understates and occasionally yields the wrong sign for private schools' impact on achievement—providing a compelling example of Lord's paradox (Lord, 1967). Whereas the restricted value-added model suggests that private schools contribute no more than public schools, our dynamic panel estimates suggest large and significant contributions ranging from 0.19 to 0.32 standard-deviations a year. From a public finance point of view, these different estimates matter particularly since per pupil expenditures are lower in private relative to public schools.[3] Our results are consistent with growing evidence that relatively inexpensive, mainstream, private schools hold potential in the developing country context (Emmanuel Jimenez and Paqueo, 1991; Alderman et al., 2001; Angrist et al., 2002; Alderman et al., 2003; Tooley and Dixon, 2003; Andrabi et al., 2008).

Our results illustrate the danger of failing to properly specify and estimate value-added models. Yet the results are not entirely negative. Despite ignoring measurement error and unobserved heterogeneity, the lagged value-added model estimated by OLS gives similar results for the private school effect as our more data intensive dynamic panel methods, although persistence remains overstated. The relative success of the lagged value-added model can be explained by the countervailing heterogeneity and measurement error biases on β and because lagged achievement can also act as a partial proxy for omitted heterogeneity in learning.[4] More generally, the bias introduced by assuming perfect persistence may not always be as severe as in our application. Both Harris and Sass (2006) and Kane and Staiger (2008), for instance, find that the persistence parameter makes little difference when estimating teacher effects. This can be explained by the small gap in baseline achievement. Children with different teachers often do not differ substantially in their baseline test scores. In contrast, given that there is little switching across school types, children currently in

[3]For details on the costs of private schooling in Pakistan see Andrabi et al. (2008).

[4]This results suggests that correcting for measurement error alone may do more harm than good. For example, Ladd and Walsh (2002) correct for measurement error in the lagged value-added model of school effects by instrumenting using double-lagged test scores but don't address potential omitted heterogeneity. They show this correction significantly changes school rankings and benefits poorly performing districts. Given that we find unobserved heterogeneity in learning rates, rankings that correct for measurement error may be poorer than those that do not.

different schools differ substantially in baseline scores. Despite this apparent robustness to different specifications when estimating teacher effects, both Kane and Staiger (2008) and Jacob et al. (2010) find that the teacher effects fade rapidly, suggesting that getting persistence right is still important to understanding long-run impacts.

The remainder of the paper is organized as follows: Section 2 presents the basic education production function analogy and discusses the specification and estimation of the value-added approximations to it. Section 3 summarizes our data. Section 4 reports our main results and several robustness checks. Section 5 concludes by discussing implications for experimental and non-experimental program evaluation.

3.2 Empirical Learning Framework

The "education production function" approach to learning relates current achievement to all previous inputs. Boardman and Murnane (1979) and Todd and Wolpin (2003) provide two accounts of this approach and the assumptions it requires; the following is a brief summary.[5] Using notation consistent with the dynamic panel literature, we aggregate all inputs into a single vector $\mathbf{x}_{it}$ and exclude interactions between past and present inputs. Achievement for child i at time (grade) t is therefore

$$y_{it}^{*} = \boldsymbol{\alpha}_1'\mathbf{x}_{it} + \boldsymbol{\alpha}_2'\mathbf{x}_{i,t-1} + \cdots + \boldsymbol{\alpha}_t'\mathbf{x}_{i1} + \sum_{s=1}^{s=t} \theta_{t+1-s}\mu_{is}, \tag{3.1}$$

where y_{it}^{*} is true achievement, measured without error, and the summed μ_{is} are cumulative productivity shocks.[6] Estimating (3.1) is generally impossible because researchers do not observe the full set of inputs, past and present. The value-added

[5] Researchers generally assume that the model is additively separable across time and that input interactions can be captured by separable linear interactions. Cunha and Heckman (2008) and Cunha et al. (2010) are two exceptions to this pattern, where dynamic complementarity between early and late investments and between cognitive and non-cognitive skills are permitted.

[6] This starting point is more restrictive than the more general starting framework presented by Todd and Wolpin (2003). In particular, it assumes an input applied in first grade has the same effect on first grade scores as an input applied in second grade has on second grade scores.

strategy makes estimation feasible by rewriting (3.1) to avoid the need for past inputs. Adding and subtracting βy^*_{it}, normalizing θ_1 to unity, and assuming that coefficients decline geometrically ($\alpha_j = \beta\alpha_{j-1}$ and $\theta_j = \beta\theta_{j-1}$ for all j) yields the *lagged value-added model*

$$y^*_{it} = \boldsymbol{\alpha}'\mathbf{x}_{it} + \beta y^*_{i,t-1} + \mu_{it}. \tag{3.2}$$

The basic idea behind this specification is that lagged achievement will capture the contribution of all previous inputs and any past unobservable endowments or shocks. As before, we refer to α as the *input coefficient* and β as the *persistence coefficient*. Finally, imposing the restriction that $\beta = 1$ yields the gain-score or *restricted value-added model* that is often used in the education literature:

$$y^*_{it} - y^*_{i,t-1} = \boldsymbol{\alpha}'\mathbf{x}_{it} + \mu_{it}.$$

This model asserts that past achievement contains no information about future gains, or equivalently, that an input's effect on any subsequent level of achievement does not depend on how long ago it was applied. As we will see from our results, the assumption that $\beta = 1$ is clearly violated in the data, and increasingly, it appears, in the literature as well. As a result, we will focus primarily on estimating (3.2).

There are two potential problems with estimating (3.2). First, the error term μ_{it} could include individual (child-level) heterogeneity in *learning* (i.e., $\mu_{it} \equiv \eta_i + \upsilon_{it}$). Lagged achievement only captures individual heterogeneity if it enters through a one-time process or endowment, but talented children may also *learn* faster. Since this unobserved heterogeneity enters in each period, $\text{Cov}(y^*_{i,t-1}, \mu_{it}) > 0$ and β will be biased upwards.

The second likely problem is that test scores are inherently a noisy measure of latent achievement. Letting $y_{it} = y^*_{it} + \varepsilon_{it}$ denote observed achievement, we can rewrite the latent lagged value-added model (3.2) in terms of observables. The full error term now includes measurement error, $\mu_{it} + \varepsilon_{it} - \beta\varepsilon_{i,t-1}$.

Dropping all the inputs to focus solely on the persistence coefficient, the expected bias due to both of these sources is

$$\operatorname{plim} \beta_{OLS} = \beta + \left(\frac{\operatorname{Cov}(\eta_i, y^*_{i,t-1})}{\sigma^2_{y^*} + \sigma^2_\varepsilon} \right) - \left(\frac{\sigma^2_\varepsilon}{\sigma^2_{y^*} + \sigma^2_\varepsilon} \right) \beta. \tag{3.3}$$

The coefficient is biased upward by learning heterogeneity and downward by measurement error. These effects only cancel exactly when $\operatorname{Cov}(\eta_i, y^*_{i,t-1}) = \sigma^2_\varepsilon \beta$ (Arellano, 2003).

Furthermore, bias in the persistence coefficient leads to bias in the input coefficients, α. To see this, consider imposing a biased $\hat{\beta}$ and estimating the resulting model

$$y_{it} - \hat{\beta} y_{i,t-1} = \boldsymbol{\alpha}' \mathbf{x}_{it} + [(\beta - \hat{\beta}) y_{i,t-1} + \mu_{it} + \varepsilon_{it} - \beta \varepsilon_{i,t-1}].$$

The error term now includes $(\beta - \hat{\beta}) y_{i,t-1}$. Since inputs and lagged achievement are generally positively correlated, the input coefficient will, in general, by biased downward if $\hat{\beta} > \beta$. The precise bias, however, depends on the degree of serial correlation of inputs and on the potential correlation between inputs and learning heterogeneity that remains in μ_{it}.

This is more clearly illustrated in the case of the restricted value-added model (assuming that $\beta = 1$) where:

$$\operatorname{plim} \hat{\alpha}_{OLS} = \alpha - (1 - \beta) \frac{\operatorname{Cov}(\mathbf{x}_{it}, y_{i,t-1})}{\operatorname{Var}(\mathbf{x}_{it})} + \frac{\operatorname{Cov}(\mathbf{x}_{it}, \eta_i)}{\operatorname{Var}(\mathbf{x}_{it})}. \tag{3.4}$$

Therefore, if indeed there is perfect persistence as assumed and if inputs are uncorrelated with η_i, OLS yields consistent estimates of the parameters α. However, if $\beta < 1$, OLS estimation of α now results in two competing biases. By assuming an incorrect persistence coefficient we leave a portion of past achievement in the error term. This misspecification biases the input coefficient downward by the first term in (3.4). The second term captures possible correlation between current inputs and

omitted learning heterogeneity. If there is none, then the second term is zero, and the bias will be unambiguously negative.

3.2.1 Addressing Child-Level Heterogeneity: Dynamic Panel Approaches to the Education Production Function

Dynamic panel approaches can address omitted child-level heterogeneity in value-added approximations of the education production function. We interpret the value-added model (3.2) as an autoregressive dynamic panel model with unobserved student-level effects:

$$y^*_{it} = \boldsymbol{\alpha}'\mathbf{x}_{it} + \beta y^*_{i,t-1} + \mu_{it}, \tag{3.5}$$

$$\mu_{it} \equiv \eta_i + \upsilon_{it}. \tag{3.6}$$

Identification of β and α is achieved by imposing appropriate moment conditions. Following Arellano and Bond (1991), we focus on linear moment conditions after differencing (3.5). In Appendix 3.A, we consider "differences and levels" GMM and "levels only" GMM, which respectively refer to whether the estimates are based on the undifferenced "levels" equation (3.5), a differenced equation (see equation (3.7) below), or both (Arellano and Bover, 1995). For more complete descriptions, Arellano and Honore (2001) and Arellano (2003) provide excellent reviews of these and other panel models.

As noted previously, the value-added model differences out omitted endowments that might be correlated with the inputs. It does not, however, difference out heterogeneity that speeds learning. To accomplish this, the basic intuition behind the Arellano and Bond (1991) difference GMM estimator is to difference again. Differencing the dynamic panel specification of the lagged value-added model (3.5) yields

$$y^*_{it} - y^*_{i,t-1} = \boldsymbol{\alpha}'(\mathbf{x}_{it} - \mathbf{x}_{i,t-1}) + \beta(y^*_{i,t-1} - y^*_{i,t-2}) + [\upsilon_{it} - \upsilon_{i,t-1}]. \tag{3.7}$$

Here, the differenced model eliminates the unobserved fixed effect η_i. However, (3.7) cannot be estimated by OLS because $y^*_{i,t-1}$ is correlated by construction with $\upsilon_{i,t-1}$

in the error term. Arellano and Bond (1991) propose instrumenting for $y^*_{i,t-1} - y^*_{i,t-2}$ using two or more period lags, such as $y^*_{i,t-2}$, or certain inputs, depending on the exogeneity conditions. These lags are uncorrelated with the error term but are correlated with the change in lagged achievement, provided $\beta < 1$. The input coefficient, in our case the added contribution of private schools, is primarily identified from the set of children who switch schools in the observation period.

The implementation of the difference GMM approach depends on the precise assumptions about inputs. We consider two candidate assumptions: strictly exogenous inputs and predetermined inputs. Strict exogeneity assumes past disturbances do not affect current and future inputs, ruling out feedback effects. In the educational context, this is a strong assumption. A child who experiences a positive or negative shock may adjust inputs in response. In our case, a shock may cause a child to switch schools.

To account for this possibility, we also consider the weaker case where inputs are predetermined but not strictly exogenous. Specifically, the predetermined inputs case assumes that inputs are uncorrelated with present and future disturbances but are potentially correlated with past disturbances. This case also assumes lagged achievement is uncorrelated with present and future disturbances. Compared to strict exogeneity, this approach uses only lagged inputs as instruments. Switching schools is instrumented by the original school type, allowing switches to depend on previous shocks. This estimator remains consistent if a child switches school at the same time as an achievement shock but still rules out parents anticipating and adjusting to future expected shocks.

Given the centrality of switchers, it is natural to consider whether the private school effect and persistence can be estimated using a differences-in-differences (DD) strategy, and how such an approach relates to our dynamic panel estimators and the differenced equation (3.7). To estimate the short-run effect α, a DD strategy could compare switchers to stayers and examine changes in test scores over two years, third and fourth grade in our data. Extending this difference-in-differences an additional year, i.e. fifth grade, gives the two year effect $\alpha(1 + \beta)$. Combined, we can recover α and β under the standard DD assumption of parallel trends.

This DD approach is related to our basic model but is not identical. The first one-year difference is similar to the differenced equation (3.7) but does not include $\beta \triangle y^*_{i,t-1}$ because $t-2$ is unavailable. The two-year differences follows by taking the two-year difference of the lagged value-added model (3.2) and expanding the terms. If we exclude fourth-to-fifth grade switchers, i.e., keep only students for whom $\mathbf{x}_{it} - \mathbf{x}_{i,t-2} = \mathbf{x}_{i,t-1} - \mathbf{x}_{i,t-2}$, the two year difference reduces to

$$\begin{aligned} y^*_{it} - y^*_{i,t-2} &= \boldsymbol{\alpha}' \left(\mathbf{x}_{it} - \mathbf{x}_{i,t-2}\right) + \beta \left(y^*_{i,t-1} - \beta y^*_{i,t-3}\right) + \mu_{it} - \mu_{i,t-2} \\ &= \boldsymbol{\alpha}' \left(\mathbf{x}_{it} - \mathbf{x}_{i,t-2}\right) + \beta \left(\boldsymbol{\alpha}' \mathbf{x}_{i,t-1} + \beta y^*_{i,t-2} + \mu_{i,t-1} - \beta y^*_{i,t-3}\right) + \mu_{it} - \mu_{i,t-2} \\ &= \boldsymbol{\alpha}'(1+\beta)\Delta \mathbf{x}_{i,t-1} + \beta \boldsymbol{\alpha}' \mathbf{x}_{i,t-2} + \left[\beta \left(\beta y^*_{i,t-2} + \mu_{i,t-1} - \beta y^*_{i,t-3}\right) + \mu_{it} - \mu_{i,t-2}\right], \end{aligned}$$

where the final equation follows from excluding fourth-to-fifth grade switchers and adding and subtracting $\beta \boldsymbol{\alpha}' \mathbf{x}_{i,t-2}$. If we focus just on private schools and incorporate the terms in the brackets into the error term, we are left with our second difference-in-differences estimate. Assuming the term in the brackets is uncorrelated with $\Delta \mathbf{x}_{i,t-1}$ and $\mathbf{x}_{i,t-2}$, the two year difference returns $\alpha(1+\beta)$.

While the DD intuition can be clarifying, our dynamic panel estimators start from the model (3.5) and estimation relies on the moment conditions explicit in the modelling; the DD approach, by comparison, makes a parallel trends assumption and estimates β indirectly. A major conclusion of this paper is that parallel trends do not imply no treatment effect if persistence is imperfect and gaps exist in baseline scores. What makes DD potentially believable it not the act of differencing but the choice of control group. The restricted value-added model, after all, can also be called a DD estimate with children in public schools forming the control group for children in private schools.

3.2.2 Addressing Measurement Error in Test Scores

Measurement error in test scores is a central feature of educational program evaluation. Ladd and Walsh (2002), Kane and Staiger (2002), and Chay et al. (2005) all

document how test-score measurement error can pose difficulties for program evaluation and value-added accountability systems. In the context of value-added estimation, measurement error attenuates the coefficient on lagged achievement and can bias the input coefficient in the process. Dynamic panel estimators do not address measurement error on their own. For instance, if we replace true achievement with observed achievement in the standard Arellano and Bond (1991) setup, (3.7) becomes

$$\triangle y_{it} = \boldsymbol{\alpha}' \triangle \mathbf{x}_{it} + \beta \triangle y_{i,t-1} + [\triangle \upsilon_{it} + \triangle \varepsilon_{i,t} - \beta \triangle \varepsilon_{i,t-1}]. \tag{3.8}$$

The standard potential instrument, $y_{i,t-2}$, is uncorrelated with $\triangle \upsilon_{it}$ but is correlated with $\triangle \varepsilon_{i,t-1} = \varepsilon_{i,t-1} - \varepsilon_{i,t-2}$ by construction.

The easiest solution is to use either three-period lagged test scores or alternate subjects as instruments. In the dynamic panel models discussed above, correcting for measurement error using additional lags requires four years of data for each child—a difficult requirement in most longitudinal datasets, including ours. We therefore use alternate subjects, although doing so does not address the possibility of correlated measurement error across subjects.e[7]

3.3 Data

To demonstrate these issues, we use data collected by the authors as part of the Learning and Educational Achievement in Punjab Schools (LEAPS) project, an ongoing survey of learning in Pakistan. The sample comprises 112 villages in 3 districts of Punjab: Attock, Faisalabad, and Rahim Yar Khan. Because the project was envisioned in part to study the dramatic rise of private schools in Pakistan, the 112 villages in these districts were chosen randomly from the list of all villages with an existing private school. As would be expected given the presence of a private school,

[7] An alternative to instrumental variables strategies is to correct for measurement error analytically using the standard error of each test score. In a working paper version of this paper we followed this strategy, using the heteroskedastic standard errors returned by Item Response Theory, and found similar results. Due to the simplicity of instrumenting using alternate subjects, we only report IV corrected estimates here.

the sample villages are generally larger, wealthier, and more educated than the average rural village. Nevertheless, at the time of the survey, more than 50 percent of the province's population resided in such villages (Andrabi et al., 2006).

The survey covers all schools within the sample village boundaries *and* within a short walk of any village household. Including schools that opened and closed over the three rounds, 858 schools were surveyed, while three refused to cooperate. Sample schools account for over 90 percent of enrollment in the sample villages.

The first panel of children consists of 13,735 third-graders, 12,110 of which were tested in Urdu, English, and mathematics. These children were subsequently followed for two years and retested in each period. Every effort was made to track children across rounds, even when they were not promoted. Nevertheless, in the tested sample, 18 percent of children were not re-tested in the second round. By the third round, 32 percent of the original tested sample is missing a fourth or fifth grade score. This is partly due to children dropping out of school (5.5 percent dropout between years 1 and 2 and another 3.2 percent between years 2 and 3) but also because of high absenteeism–just under 10 percent of children tested in the first year are absent on the day of the test in years 2 and 3. Attrition in private schools is two percentage points higher than in public schools. Children who drop out between rounds one and two have scores roughly 0.2 s.d. lower than children that don't. Controlling for school type and drop out status, drop outs in private schools are slightly better (0.05 sd) than children in public schools, although the difference is only statistically significant for math. It is plausible that the small relative differences in attrition between public and private schools imply that additional corrections for attrition are unlikely to significantly affect our results. Indeed, we explore formal corrections for attrition in Section 4.3 and find no significant changes.

In addition to being tested, 6,379 children—up to ten in each school—were randomly administered a survey including anthropometrics (height and weight) and detailed family characteristics such parental education and wealth, as measured by principal components analysis of 20 assets. When exploring the economic interpretation of persistence, we also use a smaller subsample of approximately 650 children that

can be matched to a detailed household survey that includes, among other things, child and parental time use and educational spending.

For our analysis, we use two subsamples of the data: all children who were tested in all three years (N=8120) and children who were tested *and* given a detailed child survey in all three years (N=4031). Table 3.1 presents the characteristics of these children split by whether they attend public or private schools. The patterns across each subsample are relatively stable. Children attending privates schools are slightly younger, have fewer elder siblings, and come from wealthier and more educated households. Years of schooling, which largely captures grade retention, are lower in private schools. Children in private schools are also less likely to have a father living at home, perhaps due to a migration or remittance effect on private school attendance.

The measures of achievement are based on exams in English, Urdu (the vernacular), and Mathematics. The tests were relatively long (over 40 questions per subject) and were designed to maximize the precision over a range of abilities in each grade. While a fraction of questions changed over the years, the content covered remained consistent, and a significant portion of questions appeared across all years. To avoid the possibility of cheating, the tests were administered directly by our project staff and not by classroom teachers. The tests were scored and equated across years by the authors using Item Response Theory so that the scale has cardinal meaning. Preserving cardinality is important for longitudinal analysis since many other transformations, such as the percent correct score or percentile rank, are bounded artificially by the transformations that describe them. By comparison, IRT scores attempt to ensure that change in one part of the distribution is equal to a change in another, in terms of the latent trait captured by the test. Children were tested in third, fourth, and fifth grades during the winter at roughly one year intervals. Because the school year ends in the early spring, the test scores gains from third to fourth grade are largely attributable to the fourth grade school.

Variable	Private School	Public School	Difference
Panel A: Full Sample			
Age	9.58	9.63	-0.04
	[1.49]	[1.35]	(0.08)
Female	0.45	0.47	-0.02
			(0.03)
English score (third grade)	0.74	-0.23	0.97***
	[0.61]	[0.94]	(0.05)
Urdu score (third grade)	0.52	-0.12	0.63***
	[0.78]	[0.98]	(0.05)
Math score (third grade)	0.39	-0.07	0.46***
	[0.81]	[1.00]	(0.05)
N	2337	5783	
Panel B: Surveyed Child Sample			
Age	9.63	9.72	-0.09
	[1.49]	[1.34]	(0.08)
Female	0.47	0.48	-0.02
			(0.03)
Years of schooling	3.39	3.75	-0.35***
	[1.57]	[1.10]	(0.08)
Weight z-score (normalized to U.S.)	-0.75	-0.64	-0.10
	[4.21]	[1.71]	(0.13)
Height z-score (normalized to U.S.)	-0.42	-0.22	-0.20
	[3.32]	[2.39]	(0.13)
Number of elder brothers	0.98	1.34	-0.36***
	[1.23]	[1.36]	(0.05)
Number of elder sisters	1.08	1.27	-0.19***
	[1.27]	[1.30]	(0.05)
Father lives at home	0.88	0.91	-0.04***
			(0.01)
Mother lives at home	0.98	0.98	0.00
			(0.01)
Father educated past elementary	0.64	0.46	0.18***
			(0.02)
Mother educated past elementary	0.36	0.18	0.18***
			(0.02)
Asset index (PCA)	0.78	-0.30	1.08***
	[1.50]	[1.68]	(0.07)
English score (third grade)	0.74	-0.24	0.99***
	[0.62]	[0.95]	(0.05)
Urdu score (third grade)	0.53	-0.14	0.67***
	[0.78]	[0.98]	(0.05)
Math score (third grade)	0.42	-0.09	0.51***
	[0.80]	[1.02]	(0.05)
N	1374	2657	

* Significant at the 10%; ** significant at the 5%; *** significant at 1%.

Table 3.1: Baseline characteristic of children in public and private schools. Cells contain means, brackets contain standard deviations, and parentheses contain standard errors. Standard errors for the private-public difference are clustered at the school level. Sample includes only those children tested (A) and surveyed (B) in all three years.

3.4 Results

3.4.1 Cross-sectional and Graphical Results

Before presenting our estimates of learning persistence and the implied private school effect, we provide some rough evidence for a significant private school effect using cross-sectional and graphical evidence. These results do not take advantage of the more sophisticated specifications above but nevertheless provide initial evidence that the value-added of private schools is large and significant.

3.4.1.1 Baseline estimates from cross-section data

Table 3.2 presents results for a cross-section regression of third grade achievement on child, household, and school characteristics. These regressions provide some initial evidence that the public-private gap is due to more than omitted variables and selection. Adding a comprehensive set of child and family controls reduces the estimated coefficient on private schools only slightly. Adding village fixed effects also does not change the coefficient, even though the R^2 increases substantially. Across all baseline specifications, the gap remains large: over 0.9 standard deviations in English, 0.5 standard deviations in Urdu, and 0.4 standard deviations in mathematics.

Besides the coefficient on school type, few controls are strongly associated with achievement. By far, the largest other effect is for females, who outperform their male peers in English and Urdu. However, even for Urdu, where the female effect is largest, the private school effect is still nearly three times as large. Height, assets, and whether the father (and for Column 3, mother) is educated past elementary school also enter the regression as positive and significant. More elder brothers and more years of schooling (i.e. being previously retained) correlates with lower achievement. Children with a mother living at home perform worse although this result is driven by an abnormal subpopulation of two percent of children with absent mothers. Overall, these results confirm mild positive selection into private schools but also suggest that, controlling for a host of other observables typically not available in other datasets

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable									
(third grade):	**English**	**English**	**English**	**Urdu**	**Urdu**	**Urdu**	**Math**	**Math**	**Math**
Private School	0.985	0.907	0.916	0.670	0.595	0.575	0.512	0.446	0.451
	(0.047)***	(0.048)***	(0.048)***	(0.049)***	(0.050)***	(0.047)***	(0.051)***	(0.053)***	(0.052)***
Age		0.004	0.015		0.013	0.013		0.033	0.048
		(0.013)	(0.012)		(0.013)	(0.012)		(0.014)**	(0.013)***
Female		0.125	0.133		0.209	0.205		-0.040	-0.057
		(0.047)***	(0.041)***		(0.046)***	(0.040)***		(0.051)	(0.043)
Years of schooling		-0.029	-0.019		-0.039	-0.028		-0.038	-0.025
		(0.013)**	(0.012)		(0.014)***	(0.014)**		(0.015)**	(0.014)*
Number of elder		-0.030	-0.035		-0.020	-0.025		-0.020	-0.023
brothers		(0.011)***	(0.010)***		(0.012)*	(0.011)**		(0.012)*	(0.011)**
Number of elder sisters		0.008	0.013		0.001	-0.001		-0.002	-0.006
		(0.011)	(0.010)		(0.012)	(0.012)		(0.013)	(0.012)
Height z-score		0.027	0.016		0.017	0.012		0.034	0.024
(normalized to U.S.)		(0.007)***	(0.006)***		(0.006)***	(0.006)**		(0.008)***	(0.007)***
Weight z-score		-0.005	-0.001		-0.004	0.001		-0.009	-0.002
(normalized to U.S.)		(0.008)	(0.006)		(0.005)	(0.005)		(0.007)	(0.006)
Asset index		0.041	0.050		0.043	0.045		0.030	0.034
		(0.012)***	(0.009)***		(0.011)***	(0.010)***		(0.011)***	(0.010)***
Mother educated past		0.048	0.062		0.014	0.011		0.023	-0.006
elementary		(0.036)	(0.031)**		(0.040)	(0.035)		(0.040)	(0.037)
Father educated past		0.061	0.066		0.062	0.049		0.069	0.053
elementary		(0.033)*	(0.028)**		(0.034)*	(0.031)		(0.035)**	(0.032)*
Mother lives at home		-0.131	-0.025		-0.174	-0.108		-0.210	-0.091
		(0.095)	(0.081)		(0.102)*	(0.092)		(0.097)**	(0.090)
Father lives at home		0.006	-0.038		0.019	0.005		-0.009	-0.026
		(0.049)	(0.044)		(0.053)	(0.048)		(0.057)	(0.051)
Survey Date		0.003	0.000		0.001	0.004		0.003	0.003
		(0.002)	(0.004)		(0.002)	(0.003)		(0.002)	(0.003)
Constant	-0.243	-49.721	-3.690	-0.137	-23.750	-59.528	-0.095	-56.196	-51.248
	(0.038)***	(38.467)	(62.432)	(0.035)***	(31.915)	(45.357)	(0.038)**	(35.415)	(50.310)
Village Fixed Effects	No	No	Yes	No	No	Yes	No	No	Yes
Observations	4031	4031	4031	4031	4031	4031	4031	4031	4031
R-squared	0.23	0.25	0.37	0.11	0.13	0.25	0.06	0.08	0.21

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 3.2: Third grade achievement and child, household, and school characteristics. Standard errors clustered at the school level. Sample includes only those children tested and surveyed in all three years.

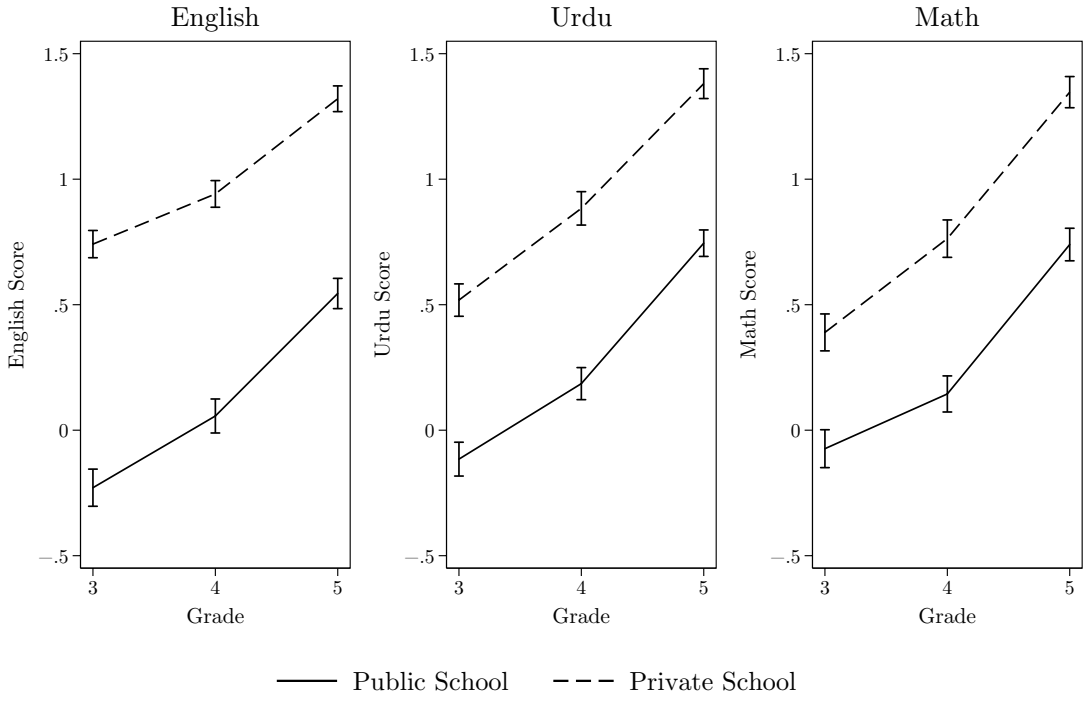
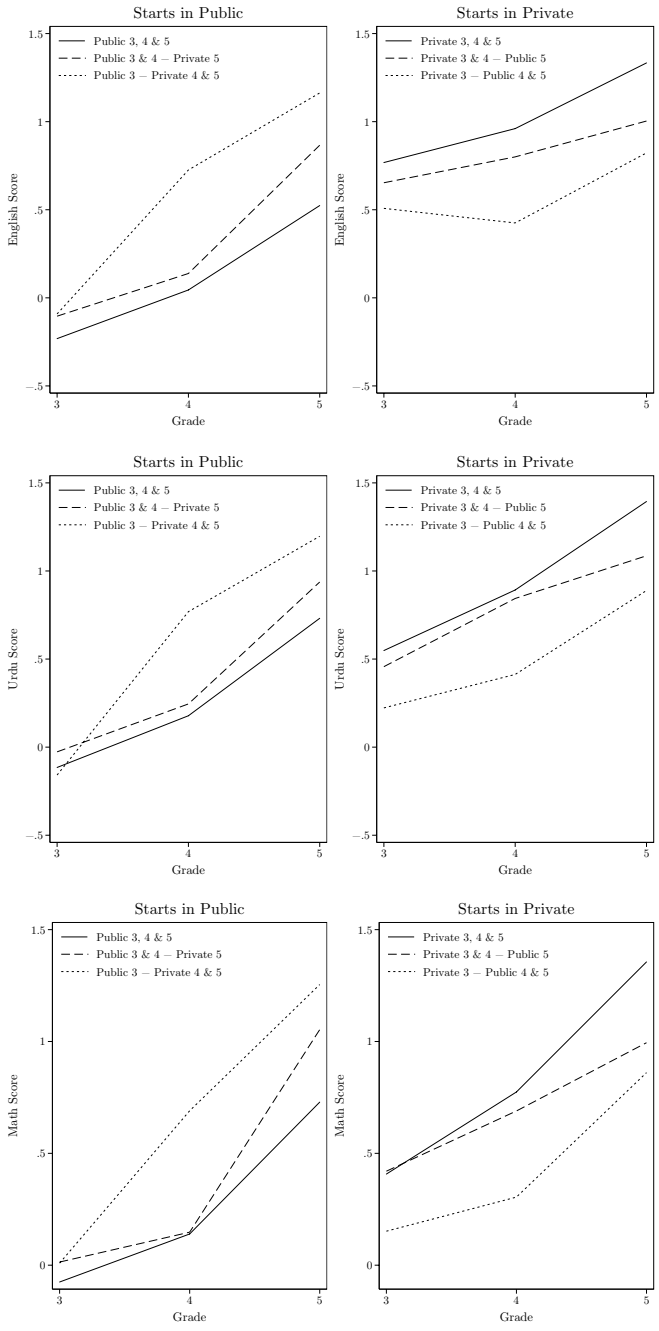


Figure 3.1: Evolution of test scores in public and private schools. Vertical bars represent 95% confidence intervals around the group means, allowing for arbitrary clustering within schools. The graph's sample is limited to children who were tested in all three periods.

(such as child height and household assets) does not alter significantly the size of the private schooling coefficient.

3.4.1.2 Graphical and reduced-form evidence

Figure 3.1 plots learning levels in the tested subjects (English, mathematics, and the vernacular, Urdu) over three years. While, levels are always higher for children in private schools, there is little difference in learning gains (the gradient) between public and private schools. This illustrates why a specification that uses learning gains (i.e., assumes perfect persistence) would conclude that private schools add no greater value to learning than their public counterparts.



	Public 3, 4, & 5	**Public 3, 4 – Private 5**	**Public 3 – Private 4 & 5**	**Private 3, 4 & 5**	**Private 3 & 4 – Public 5**	**Private 3 – Public 4, 5**
N	5688	40	48	2007	160	167

Figure 3.2: Achievement over time for children who switched school types. Lines connect group means for children who were enrolled in all three periods and have a particular private/public enrollment pattern. Children were tested in the second half of the school year; most of the gains from a child in a third grade government school and fourth grade private school should be attributed to the private school.

The dynamic panel estimators that we explore identify the private school effect using children who switch schools. Figure 3.2 illustrates the patterns of achievement for these children. For each subject we plot two panels: the first containing children who start in public school and the second containing those who start in private school. We then graph achievement patterns for children who never switch, switch after third grade, and switch after fourth grade. For simplicity, we exclude children who switch back and forth between school types.

As the table at the bottom of the figure shows, very few children change schools. Only 48 children move from public to private schools in fourth grade, while 40 move in fifth grade. Consistent with the role of private schools serving primarily younger children, 167 children switch to public schools in fourth grade, and 160 switch in fifth grade. These numbers are roughly double the number of children available for our estimates that include controls, since only a random subset of children were surveyed regarding their family characteristics.

Even given the small number of children switching school types, Figure 3.2 provides preliminary evidence that the private school effect is not simply a cross-sectional phenomenon. In all three subjects, children who switch to private schools between third and fourth grade experience large achievement gains. Children switching from private schools to public schools exhibit similar achievement patterns, except reversed. Moving to a public school is associated with slower learning or even learning losses. Most gains or losses occur immediately after moving; once achievement converges to the new level, children experience parallel growth in public and private schools.

These results are consistent with low persistence and a large private school effect. Consider, for instance, the panel for Urdu and children starting in public schools (middle, left). Children who switch to private schools in fourth grade experience large immediate gains compared to children that stay in public schools. A difference-in-differences analysis would therefore indicate a large private school effect α. However, if we extend this difference-in-differences an additional year to include fifth grade, the estimate remains virtually unchanged. That is, Figure 2 suggests the two year effect is roughly the same as the one year effect, or, equivalently, that $\alpha \approx \alpha(1 + \beta)$. If $\alpha > 0$, this is only possible if $\beta \approx 0$.

	One and two-year treatment effect		
Length of treatment	**English**	**Urdu**	**Math**
One year (gain between third and fourth)	0.31***	0.26***	0.26***
	(0.05)	(0.07)	(0.06)
Two years (gain between third and fifth)	0.33***	0.28***	0.31***
	(0.05)	(0.05)	(0.06)

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 3.3: Differences-in-differences estimates of short- and long-run private school effect. Standard errors clustered at the school level. The coefficients are from a regression of the change in scores between third and fourth grade (first line) and between third and fifth grade (second line) on third grade school type, and switching (treatment), pooled by defining switchers as 1 for public-private, 0 for public-public and private-private, and -1 for private-public. The sample excludes children who switched between fourth and fifth grade, making students that stay in public or private school the comparison group for students switching between third and fourth grade.

Table 3.3 confirms this difference-in-differences intuition. We regress changes in achievement between third and fourth grade (column one) and between third and fifth grade (column two) on third grade school type and switching between third and fourth grade (the treatment). We pool switchers so that -1 denotes switchers from private to public, 0 denotes stayers, and 1 denotes switchers from public to private, and exclude children that switch between fourth and fifth grade. Students that stay in public or private schools therefore form the comparison group for students that switched between third and fourth grade. As Table 3.3 shows, the estimated two year treatment effect $\alpha(1+\beta)$ is only slightly higher than the estimated one year treatment effect α, consistent with low persistence. Our subsequent results, which use the full dynamic panel setup, yield similar estimates to this simpler difference-in-differences approach.

The results in Table 3.3 and dynamic panel estimators rely on children that switch school types. A potential concern is that children who switch schools are more likely to have experienced changes in their family circumstances during the year. To the extent

that changes in household circumstances impact family's investment in children, our coefficients could be biased.

To examine this issue further, we examine the correlation between switching to or from a private school—again defined as 1, 0, and -1—and changes in the family's assets and wealth, the presence of parents, and child height, weight, and health. Our dynamic panel estimates control for these observable changes, but large comovements with school switching would be worrying. Table 3.4 reports both changes in characteristics for future and contemporaneous switchers. We include three samples: : surveyed children (Table 3.1, Panel B), surveyed children matched to a household survey, and any child found in the household survey. We include this final group, which includes all grades, to increase the sample size for the child health and household asset questions.

Across these three samples, and for both pre-trend and contemporaneous switching, we find that household characteristics do not comove with school switches in a direction that would favor private schools. The only large and statistically significant correlation is a *negative* correlation between contemporaneous switching to a private school and child height and weight. These coefficients are of the order of 0.2 standard deviations and significant at the 1 percent (weight) and 5 percent (height) level. Child health is also negatively correlated with switching to a private school, although is not statistically significant. The correlations with child health and anthropometrics are puzzling. One possibility is that households compensated for greater educational investments in children (enrolling them in private school) by *reducing* their investments in health. To the extent that this is a causal impact, it suggests that the benefits of private schooling are somewhat *reduced* due to household compensations on other dimensions, in particular, child nutrition. The results that follow are essentially a more careful analysis that includes the possibility of unobserved heterogeneity and corrects for measurement error, both of which we find are central complications in value-added models.

Changes in Time-Varying Child Characteristics	**Contemporaneous Switcher**	**Future Switcher**
Weight z-score	-.25***	0.11
	(0.07)	(0.09)
Height z-score	-.19**	0.15
	(0.09)	(0.12)
Asset index (PCA)	.13	-0.05
	(0.08)	(0.09)
Mother lives at home	-0.00	-0.03*
	(0.01)	(0.02)
Father lives at home	0.00	-0.04
	(0.02)	(0.03)
Relative wealth (household survey)	2.08	3.04
	(3.77)	(6.02)
Asset index (household survey)	0.22	-0.57
	(0.40)	(0.68)
Child health (household survey)	-0.40	-0.17
	(0.27)	(0.34)
Child health (household survey, all grades)	-0.14	(0.05)
	(0.09)	(0.08)
Relative wealth (household survey, all grades)	-1.33	2.20
	(2.11)	(1.83)

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 3.4: School type switchers and time-varying child characteristics. Numbers are coefficients from a regression of changes in time-varying child characteristics (dependent variable) on pooled school switching (defined as public-private = 1, public-public or private-private = 0, private-public = -1). Contemporaneous switchers use changes in characteristics and school type in the same period, whereas future switchers compares switching with changes in the preceding year. Standard errors are clustered at the school level. Household survey variables include matched children (household survey) and, separately, children switching in any grade (household survey, all grades). Height and weight -z-scores are normalized to the US average; asset indices are from a PCA analysis of household assets; relative wealth is a subjective question asked so that 100 represents the average village wealth; health is measured on a scale from 1 (bad) to 16 (perfect) health.

Model (Key Assumption, Estimator)	English			Urdu			Math		
	Persistence	Private School	Hansen's J	Persistence	Private School	Hansen's J	Persistence	Private School	Hansen's J
M1. Restricted value-added (perfect persistence	1.00	-0.08		1.00	0.01		1.00	0.05	
β=1, OLS)		(0.02)			(0.02)			(0.02)	
M2. Lagged value-added (no effects, no	0.52	0.31		0.58	0.26		0.57	0.27	
measurement error, OLS)	(0.02)	(0.02)		(0.01)	(0.02)		(0.02)	(0.03)	
M3. Lagged value-added with measurement error	0.70	0.16	4.69	0.73	0.17	3.67	0.76	0.17	0.02
correction (no effects, 2SLS)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.06)	(0.02)	(0.03)	(0.89)
M4. Differenced dynamic panel, strictly exogenous	0.19	0.25	25.44	0.21	0.29	49.50	-0.00	0.26	33.97
inputs (GMM)	(0.10)	(0.07)	(0.02)	(0.09)	(0.07)	(0.00)	(0.09)	(0.08)	(0.00)
M5. Differenced dynamic panel, predetermined	0.19	1.15	16.82	0.35	0.90	18.90	0.12	0.46	12.06
inputs (GMM)	(0.10)	(0.39)	(0.02)	(0.11)	(0.48)	(0.01)	(0.12)	(0.50)	(0.10)

Notes: Cells contain estimates for the key parameters and standard errors clustered by school. M3 corrects for measurement error using alternate subjects. M4 and M5 use instruments use twice lagged alternate scores and differenced (strictly exogenous) or lagged (predetermined) covariates. Hansen's J reports the chi2 and associated p-value with df=1, 13 and 7 for models M3-M5.

Table 3.5: Restricted value-added, lagged value-added, lagged value-added with measurement error correction, and differenced dynamic panel models.

3.4.2 OLS and Dynamic Panel Value-Added Estimates

Tables 3.5 summarize our main value-added results. All estimates include the full set of controls in the child survey sample, the survey date, round (grade) dummies, and village fixed effects. For brevity, we only report the persistence and private school coefficients.[8] We group the discussion of our results in three domains: estimates of the persistence coefficient, estimates of the private schooling coefficient, and regression diagnostics.

3.4.2.1 The persistence parameter

We immediately reject the hypothesis of perfect persistence ($\beta = 1$). Across all specifications and all subjects (except M1 which imposes $\beta = 1$), the estimated persistence coefficient is significantly lower than one, even in the specifications that correct for measurement error only and should be biased upward (M3 and M4). The typical lagged value-added model (M2), which assumes no omitted student heterogeneity and no measurement error, returns estimates between 0.52 and 0.58 for the persistence coefficient. Correcting only for measurement error by instrumenting using the two alternate subjects (M3) increases the persistence coefficient to between 0.70 and 0.79,

[8] As discussed, time-invariant controls drop out of the differenced models. For the system and levels estimators reported in Appendix 3.A we also assume, by necessity, that time-invariant controls are uncorrelated with the fixed effect or act as proxy variables.

consistent with significant measurement error attenuation. This estimate, however, remains biased upward by omitted heterogeneity.

Moving to our dynamic panel estimators, Table 3.5 gives the Arellano and Bond (1991) difference GMM estimates under the assumption that inputs are strictly exogenous (M4) or predetermined (M5). In English and Urdu, the persistence parameter falls to between 0.19 and 0.35. The estimates are (statistically) different from models that correct for measurement error only. *In other words, omitted heterogeneity in learning exists, and biases the static estimates upward.* For mathematics, the estimated persistence coefficient is indistinguishable from zero, considerably below all the other estimates. These estimates are higher and somewhat more stable in the systems GMM approach summarized in Appendix 3.A.

3.4.2.2 The contribution of private schools

Assuming perfect persistence biases the private school coefficient downward. For English, the estimated private school effect in the restricted model that incorrectly assumes $\beta = 1$ is negative and significant. For Urdu and mathematics, the private school coefficient is small and insignificant or marginally significant. By comparison, the dynamic panel estimates are positive and statistically significant, with the exception of one of the predetermined difference GMM estimates, which is too weak to identify the private school effect with any precision. The additional estimators in Appendix A follow a similar pattern.

An overarching theme in this analysis is that the persistence parameter influences the estimated private school effect, but that low precision makes it difficult to distinguish estimates based on different exogeneity conditions. This is largely due to the small number of children switching between public and private schools in our sample. Rather than estimating the persistence coefficient, we could assume a specific rate and then estimate the value-added model. That is, we use $y_{it} - \beta y_{i,t-1}$ as the dependent variable. This provides a robustness check for any estimated effects, requires only two years of data, and eliminates the need for complicated measurement error corrections. (It assumes, however, that inputs are uncorrelated with the omitted learning heterogeneity.) Moving from the restricted value-added model ($\beta = 1$) to

the pooled cross-section model ($\beta = 0$) increases the estimated effect from negative or insignificant to large and significant. For most of the range of the persistence parameter, the private school effect is positive and significant, but pinning down the precise yearly contribution of private schooling depends on our assumptions about how children learn.

A couple of natural questions are how these estimates compare to the private-public differences reported in the cross-section and why the trajectories in Figure 1 are parallel even though the private school effect is positive. Controlling for observables suggests that, after three years, children in private schools are 0.9 (English), 0.5 (Urdu), and 0.45 (mathematics) standard deviations ahead of their public school counterparts. If persistence is 0.4 and the yearly private school effect is 0.3, children's trajectories will become parallel when that achievement gap reaches 0.5 ($= 0.3/(1 - 0.4)$). This is roughly the gap we find in Urdu and mathematics. Any small disagreement, including the larger gap in English, may be attributable to baseline selection effects. Thus, our results can consistently explain the large baseline gap in achievement, the parallel achievement trajectories in public and private schools, and the significant and ongoing positive private school effect.

3.4.2.3 Regression diagnostics

For many of the GMM estimates, Hansen's J test rejects the overidentifying restrictions implied by the model. This is troubling but not entirely unexpected. Different instruments may be identifying different local average treatment effects in the education context. For example, the portion of third grade achievement that remains correlated with fourth grade achievement may decay at a different rate than what was learned most recently. This is particularly true in an optimizing model of skill formation where parents smooth away shocks to achievement. In such a model, unexpected shocks to achievement, beyond measurement error, would fade more quickly than expected gains. Instrumenting using contemporaneous alternate subject scores will therefore more likely identify different parameters than instrumenting using previous year scores. Likewise, instrumenting using alternate lags, differenced achievement and/or inputs may also identify different effects. One result of note is that dropping

the overidentifying inputs typically raises the the persistence coefficient slightly, to roughly 0.25 for math. This type of heterogeneity is important and suggests that a richer model than a constant coefficient lagged value-added may be warranted. [9]

3.5 Conclusion

In the absence of randomized studies, the value-added approach to estimating education production functions has gained momentum as a valid methodology for removing unobserved individual heterogeneity in assessing the contribution of specific programs or in understanding the contribution of school-level factors for learning (e.g. Boardman and Murnane, 1979; Hanushek, 1979; Todd and Wolpin, 2003; Hanushek, 2003; Doran and Izumi, 2004; McCaffrey, 2004; Gordon et al., 2006). In such models, assumptions about learning persistence and unobserved heterogeneity play central roles. Our results reject both the assumption of perfect persistence required for the restricted value-added model and of no learning heterogeneity required for the lagged value-added model. Our results for Pakistan should illustrate the danger of incorrectly modeling or estimating education production functions: the restricted value-added model is fundamentally misspecified and can even yield wrong-signed estimates of a program's impact. Underscoring the potential of affordable, mainstream private schools in developing countries, we find that Pakistan's private schools contribute roughly 0.25 standard deviations more to achievement each year than government schools, an effect greater than the average yearly gain between third and fourth grade.

Our estimates of low persistence are consistent with recent work on teacher effects and with experimental evidence of program fade out in developing and developed countries. Table 3.6 summarizes eight randomized (or quasi-randomized) interventions that followed children after the program ended. This follow-up enables estimation of both immediate and extended treatment effects. For the interventions summarized, the extended treatment effect represents test scores roughly one year

[9] In a working paper version of this paper, we performed a series of additional robustness checks to test the plausibility of low persistence. The rates we find are consistent other work in the literature, with the expected bias under plausible parametrization of the lagged value-added model, and with the expected bias under the assumption of equal selection on observables and unobservables.

Program	Subject	Immediate Treatment Effect	Extended Treatment Effect	Implied Persistence Coefficient	Source
Balsakhi Program	Math	0.348	0.030	0.086	Banerjee et al (2007)
	Verbal	0.227	0.014	0.062	
CAL Program	Math	0.366	0.097	0.265	Banerjee et al (2007)
	Verbal	0.014	-0.078	~0.0	
Learning Incentives	Multi-subject	0.23	0.16	0.70	Kremer et al (2003)
Teacher Incentives	Multi-subject	0.139	-0.008	~0.0	Glewwe et al (2003)
Tracked Classes	Multi-subject	0.138	0.163	1.2	Duflo, Dupas, and Kremer(2009)
Contract Teachers	Multi-subject	0.181	0.094	0.52	Duflo, Dupas, and Kremer(2009)
STAR Class Size Experiment	Stanford-9 and CTBS	~5 percentile points	~2 percentile points	~ .25 to .5	Krueger and Whitmore (2001)
Summer School and Grade Retention	Math	0.136	0.095	0.70	Jacob and Lefgren (2004)
	Reading	0.104	0.062	0.60	

Table 3.6: Experimental estimates of program fade out. Extended treatment effect is achievement approximately one year after the treatment ended. Unless otherwise noted, effects are expressed in standard deviations. Results for Kremer et al. (2003) are averaged across boys and girls. Estimated effects for Jacob and Lefgren (2004) are taken for the third grade sample.

after the particular program ended. For a number of the interventions, the persistence coefficient is less than 0.10. In two interventions—learning incentives and grade retention—the coefficient is between 0.6 and 0.7. However, this higher level of persistence may in part be explained by the specific nature of these interventions.[10] Perhaps most interestingly, Duflo et al. (2010) report results from an experiment providing both additional contract teachers and tracking students. While the impact of contract teachers fades out, consistent with our , the effect of thetracking treatment increases over time in the same experimental context, even though children returned to the same classes after the experiment concluded. Although the link between fade out in experimental studies and the persistence parameter is not always exact, the evidence from several randomized studies suggests that current learning does not always

[10]In the case of grade retention, there is no real "post treatment" period since children always remain one grade behind after being retained. If one views grade retention as an ongoing multi-period treatment, then lasting effects can be consistent with low persistence. In the case of learning incentives, Kremer et al. (2003) argue that student incentives increased effort (not just achievement) even after the program ended, leading to ongoing learning.

carry over to future learning without loss, and in fact, these losses may be substantial for most treatments. The results of Duflo et al. (2010) suggest that evaluating long-run outcomes is critical to understanding the ultimate efficacy of educational interventions.

But the economic interpretation of low persistence still remains area open to enquiry. Our context and test largely rule out mechanical explanations of low persistence such as psychometric bounding effects, cheating, or changing content. In a preliminary exploration reported in a working paper version of this paper, we also found little evidence that low persistence results from substitution by parents and teachers. Simple forgetting, consistent with a large body of memory research in psychology, appears to be a likely explanation and hence a core component of education production functions. But more research is needed to provide direct evidence for it, and to understand whether the inability to perform on a test implies that the underlying knowledge has been truly lost.

Our results also suggest that short evaluations, even when experimental, may yield little information about the cost-effectiveness of a program. Using the one or two year increase from a program gives an upper-bound on the longer term achievement gains. As our estimates suggest, and Table 6 confirms, we should expect program impacts to fade quickly. Calculating the internal rate of return by citing research linking test scores to earnings of young adults is therefore a doubtful proposition. The techniques described here, with three periods of data, can theoretically obtain a lower bound on cost-effectiveness by assuming exponential fade out. At the same time, the causes of fade out are equally important: if parents no longer need to hire tutors or buy textbooks (the substitution interpretation of imperfect persistence), a program may be *cost*-effective even if test scores fade out.

Moving forward, empirical estimates of education production functions may benefit from further unpacking persistence. Overall, the agenda pleads for a richer model of education and for empirical techniques for modelling the broader learning process, not simply to add nuance to our understanding of learning, but to get the most basic parameters right.

3.A Additional Estimation Strategies

3.A.1 System GMM

3.A.1.1 Uncorrelated or constantly correlated effects

One difficulty with the differences GMM approach (M4 and M5) is that time-invariant inputs drop out of the estimated equation and their effects are therefore not identified. In our case, this means that the identification of the private school effect is based on the five percent of children who switch between public and private schools. This leads to large standard errors in Table 3.5. We address the limited time-series variation using the levels and differences GMM framework proposed by Arellano and Bover (1995) and extended by Blundell and Bond (1998). Levels and differences GMM estimates a system of equations, one for the undifferenced levels equation (3.5) and another for the differenced equation (3.7). Further assumptions regarding the correlation between inputs and heterogeneity (though not necessarily between heterogeneity and lagged achievement) yield additional instruments.

We first consider predetermined inputs that have a constant correlation with the individual effects (M6). While inputs may be correlated with the omitted effects, constant correlation implies switching is not. The constant correlation assumption implies that $\triangle \mathbf{x}_{it}$ are available as instruments in the levels equation (Arellano and Bover, 1995). In the context of estimating school type, this estimator can be viewed as a levels and differences switching estimator since it relies on children switching school types in both the levels and differences equations. In practice, we often must assume that any time-invariant inputs are uncorrelated with the fixed effect or the levels equation, which includes the time-invariant inputs, is not fully identified.

A second possibility is that inputs are predetermined but are also uncorrelated with the omitted effects (M7). This allows using inputs $\mathbf{x}_i^t$ as instruments in the levels model (3.5). The required assumption is fairly strong; it is natural to believe that inputs are correlated with the omitted effect. Certainly, the decision to attend private school may be correlated with the child's ability to learn although a rich enough set of controls may make the assumption plausible. The assumption is weaker than OLS

estimation of lagged value-added model since the model (M7) allows for the omitted fixed effect to be correlated with lagged achievement.

3.A.1.2 Conditional mean stationarity

In some instances, it may be reasonable to assume that, while learning heterogeneity exists, it does not affect achievement gains. A talented child may be so far ahead that imperfect persistence cancels the benefit of faster learning. That is, individual heterogeneity may be uncorrelated with gains, $y^*_{it} - y^*_{it-1}$, but not necessarily with *learning,* $y^*_{it} - \beta y^*_{it-1}$. This situation arises when the initial conditions have reached a convergent level with respect to the fixed effect such that

$$y^*_{i1} = \frac{\eta_i}{1-\beta} + d_i, \tag{3.9}$$

where $t = 1$ is the first observed period and not the first period in the learning life-cycle. Blundell and Bond (1998) discuss this type of conditional mean stationarity restriction in considerable depth. As they point out, the key assumption is that initial deviations, d_i, are uncorrelated with the level of $\eta_i/(1-\beta)$. It does not imply that the achievement path, $\{y^*_{i1}, y^*_{i2}, \ldots, y^*_{iT}\}$, is stationary; inputs, including time dummies, continue to spur achievement and can be nonstationary. The assumption only requires that, conditional on the full set of controls and common time dummies, the individual effect does not influence achievement gains.

While this assumption seems too strong in the context of education, we discuss it because the dynamic panel literature has documented large downward biases of other estimators when the instruments are weak (e.g. Blundell and Bond, 1998). This occurs when persistence is perfect ($\beta = 1$) since the lagged value-added model then exhibits a unit root and lagged test scores become weak instruments in the differenced model. The conditional mean stationarity assumption provides an additional $T - 2$ non-redundant moment conditions that can augment the system GMM estimators. While a fully efficient approach uses these additional moments along with typical moments in the differenced equation, the conditional mean stationarity assumption ensures strong instruments in the levels equation to identify β. Thus, if we prefer

Model (Key Assumption, Estimator)	English			Urdu			Math		
	Persistence	Private School	Hansen's J	Persistence	Private School	Hansen's J	Persistence	Private School	Hansen's J
Levels and Difference SGMM									
M6. Predetermined inputs, constantly correlated effects	0.36	0.21	45.50	0.26	0.22	66.58	0.12	0.19	57.63
	(0.07)	(0.06)	0.00	(0.08)	(0.06)	(0.00)	(0.10)	(0.08)	(0.00)
M7. Predetermined inputs, uncorrelated effects	0.53	0.32	79.08	0.51	0.30	81.89	0.51	0.30	82.19
	(0.05)	(0.04)	0.00	(0.06)	(0.04)	(0.00)	(0.08)	(0.05)	(0.00)
Levels Only GMM									
M8. Predetermined inputs, constantly correlated effects, conditional stationarity	0.40	0.29	24.74	0.55	0.31	13.49	0.51	0.30	29.45
	(0.05)	(0.07)	(0.02)	(0.05)	(0.07)	(0.33)	(0.06)	(0.07)	(0.00)
M9. Predetermined inputs, uncorrelated effects, conditional stationarity	0.39	0.24	23.43	0.56	0.27	13.30	0.53	0.27	28.36
	(0.05)	(0.04)	(0.02)	(0.05)	(0.03)	(0.27)	(0.06)	(0.04)	(0.00)

Table 3.7: System and levels-only dynamic panel models. Cells contain estimates for the key parameters and standard errors clustered by school. M6 and M7 are a system estimators, including both a difference and levels equation with differenced (M6) or undifferenced (M7) covariates as additional instruments in the levels equation.. M8 and M9 use only the levels equation for simplicity and included differenced scores as an additional instrument. Hansen's J reports the chi2 and associated p-value with df=23, 29, 12, and 11.

simplicity over efficiency, we can estimate the model using levels GMM or 2SLS and avoid the need to use a system estimator. In this simpler approach, we instrument the undifferenced value-added model (3.5) using lagged changes in achievement, $\triangle y_i^{*t-1}$, and either changes in inputs, $\triangle \mathbf{x}_i^t$, or inputs directly, $\mathbf{x}_i^t$, depending on whether inputs are constantly correlated (M8) or are uncorrelated with the individual effect (M9).

3.A.1.3 Results

Table 3.7 reports the persistence and private school effect for these additional estimators. Most estimators have higher, but still lower the the lagged model, persistence estimates. With the addition of a conditional mean stationarity assumption, we can more precisely estimate the persistence coefficient. In this model, we only use moments in levels to illustrate a dynamic panel estimator that improves over the lagged value-added model estimated by OLS but doesn't require estimating a system of equations. The persistence coefficient rises substantially to between 0.39 and 0.56. This upward movement is consistent with a violation of the stationarity assumption (the fixed-effect still contributes to achievement growth) but an overall reduction in the omitted heterogeneity bias. Across the various dynamic panel models and subjects, estimates of the persistence parameter vary from 0.2 to 0.55. However, the highest

dynamic panel estimates come from assuming conditional mean stationary, which is likely too strong an assumption in the context of learning.

Adding a levels equation and using the assumption that inputs are constantly correlated or uncorrelated with the omitted effects reduces the standard errors for the private school coefficient while maintaining the assumption that inputs are predetermined but not strictly exogenous. Under the scenario that private school enrollment is constantly correlated with the omitted effect (M6), the private school coefficient is large: 0.19 to 0.32 standard deviations (depending on the subject) and statistically significant. This estimate allows for past achievement shocks to affect enrollment decisions but assumes that switching school type is uncorrelated with unobserved student heterogeneity. Within the systems context, this is our preferred estimate.

3.A.2 Attrition corrected estimators

One potential explanation for low persistence is attrition. Roughly a third of our original tested sample cannot be included in our estimates due to missing intermediate or final test scores. Lower scoring students are more likely to attrit and it is possible that these students also experience little growth in learning from year to year. If so, these students will display high persistence in test scores year to year and excluding them from the analysis will bias our estimates downward. We find little evidence that this is a significant source of bias. Using the sample of children that attrit in fifth grade yields similar or even slightly lower estimates for persistence.

To fully correct for attrition in moment-based model such as ours, Abowd et al. (2001) propose weighting by the estimated inverse probability that each observation remains in the sample. Analogous to propensity score weighting in the program evaluation literature, inverse probability weighting eliminates potential attrition bias if attrition is based on observables. To evaluate potential attrition bias, we estimate the probability of attrition using all past test scores and child characteristics and report our weighted results for two models in Table 3.8. Reassuringly, these corrections for attrition make little difference. Both the private school coefficient and the persistence coefficient change only slightly compared to Tables 3.5 and 3.7, and the direction of

Strategy	**English**	**Urdu**	**Math**
M2. No effects, no measurement error (OLS)			
Private school	0.28	0.22	0.22
	(0.03)	(0.03)	(0.03)
Persistence	0.55	0.59	0.65
	(0.02)	(0.02)	(0.02)
M8: Predetermined inputs, constantly correlated effects, conditional stationarity			
Private school	0.25	0.28	0.30
	(0.07)	(0.08)	(0.09)
Persistence	0.40	0.55	0.50
	(0.05)	(0.05)	(0.06)

Table 3.8: Correcting for potential attrition bias with inverse probability weighting. Cells give model coefficients and parenthesis give standard errors. Model mirrors the uncorrected estimates but weight the sample using the estimated inverse probability of an observation being observed. The probability of attrition is estimated from two logit regressions for fourth and fifth grade including all previous data as controls and village fixed effects. See Abowd, Crepon and Kramarz (2001) for details.

change differs across models and subjects. The likely explanation for why corrections for attrition do not affect our estimates is that the bulk of children who are not tested in any given year are not drop-outs but children absent on the day of the test, which may be a largely random process. Indeed, simple OLS estimates of persistence based on the sub-sample of children who report only 2 years of test-scores are within 0.05 standard deviations of estimates based on children who were present for all 3 tests, and the difference is statistically insignificant.

Bibliography

A. Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

J. Abbring and J. Heckman. Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. *Handbook of Econometrics*, 6:5145–5303, 2007.

J. M. Abowd, B. Crepon, and F. Kramarz. Moment Estimation with Attrition: An Application to Economic Models. *Journal of the American Statistical Association*, 96(456):1223–1232, 2001.

H. Alderman, P. F. Orazem, and E. M. Paterno. School quality, school cost, and the public/private school choices of low-income households in pakistan. *The Journal of Human Resources*, 36(2):304–326, 2001.

H. Alderman, J. Kim, and P. F. Orazem. Design, Evaluation, and Sustainability of Private Schools for the Poor: The Pakistan Urban and Rural Fellowship School Experiments. *Economics of Education Review*, 22(3):265–274, 2003.

T. Andrabi, J. Das, and A. I. Khwaja. A dime a day : the possibilities and limits of private schooling in pakistan. *World Bank Policy Research Working Paper 4066*, November 2006.

T. Andrabi, J. Das, and A. I. Khwaja. A dime a day: The possibilities and limits of private schooling in Pakistan. *Comparative Education Review*, 52(3):329–355, 2008.

J. D. Angrist, E. P. Bettinger, E. Bloom, E. King, and M. Kremer. Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment. *The American Economic Review*, 92(5):1535–1558, 2002.

M. Arellano. *Panel Data Econometrics*. Oxford University Press, 2003.

M. Arellano and S. Bond. Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, 58(2):277–297, 1991.

M. Arellano and O. Bover. Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, 68(1):29–51, 1995.

M. Arellano and B. E. Honore. Panel data models: some recent developments. In J. J. Heckman and E. E. Leamer, editors, *Handbook of Econometrics*, volume 5 of *Handbook of Econometrics*, chapter 53, pages 3229–3296. Elsevier, 03 2001.

A. V. Banerjee, S. A. Cole, E. Duflo, and L. L. Linden. Remedying Education: Evidence from Two Randomized Experiments in India. *Quarterly Journal of Economics*, 122(3), August 2007.

J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

S. Black. Do Better Schools Matter? Parental Valuation of Elementary Education. *The Quarterly Journal of Economics*, 114(2):577–599, 1999.

R. Blundell and S. Bond. Initial conditions and Moment Conditions in Dynamic Panel Data Models. *Journal of Econometrics*, 87(1):115–43, 1998.

A. E. Boardman and R. J. Murnane. Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education*, 52(2):113–121, 1979.

G. Brunello and D. Checchi. Does school tracking affect equality of opportunity? New international evidence. *Economic Policy*, 22(52):781–861, 2007.

K. Y. Chay, P. J. McEwan, and M. S. Urquiola. The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools. *The American Economic Review*, 95(4):1237–1258, 2005.

M. Cheng, J. Fan, and J. Marron. On Automatic Boundary Corrections. *The Annals of Statistics*, 25(4):1691–1708, 1997.

T. Cook. "Waiting for Life to Arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics. *Journal of Econometrics*, 142(2):636–654, 2008.

F. Cunha and J. J. Heckman. Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Journal of Human Resources*, 43 (4):738–782, 2008.

F. Cunha, J. J. Heckman, and S. M. Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883, May 2010.

J. Currie and D. Thomas. Does Head Start Make a Difference? *The American Economic Review*, 85(3):341–364, 1995.

I. Davidoff and A. Leigh. How Much do Public Schools Really Cost? Estimating the Relationship between House Prices and School Quality*. *Economic Record*, 84 (265):193–206, 2008.

A. Davison and D. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, 1997.

R. Dehejia. Program evaluation as a decision problem. *Journal of Econometrics*, 125 (1-2):141–173, 2005.

M. Dell. The Persistent Effects of Peru's Mining Mita. *Working Paper*, 2010.

D. Deming. Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics*, 1(3): 111–134, 2009.

H. Doran and L. T. Izumi. Putting Education to the Test: A Value-Added Model for California. *San Francisco: Pacific Research Institute*, 2004.

E. Duflo, P. Dupas, and M. Kremer. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *NBER Working Paper No. 14475*, 2010.

Y. Edwards and G. Allenby. Multivariate analysis of multiple response data. *Journal of Marketing Research*, 40(3):321–334, 2003.

M. E. L. Emmanuel Jimenez and V. Paqueo. The relative efficiency of private and public schools in developing countries. *The World Bank Research Observer*, 6(2): 205–218, 1991.

J. Fan. Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, pages 998–1004, 1992.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. CRC Press, 1996.

B. Frandsen and R. Guiteras. Using regression discontinuity to estimate the distributional effects of educational interventions. *Mimeo*, 2010.

C. Frangakis and D. Rubin. Principal Stratification in Causal Inference. *Biometrics*, 58(1):21–29, 2002.

P. Fredriksson and B. Holmlund. Improving incentives in unemployment insurance: A review of recent research. *Journal of Economic Surveys*, 20(3):357, 2006.

M. Frölich. Regression discontinuity design with covariates. IZA Discussion Papers 3024, Institute for the Study of Labor (IZA), Sept. 2007.

A. Gelman, X. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–759, 1996.

A. Gelman, I. Van Mechelen, G. Verbeke, D. Heitjan, and M. Meulders. Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics*, 61(1):74–85, 2005.

A. Gelman, A. Jakulin, M. Pittau, and Y. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

A. S. Gerber, D. P. Kessler, and M. Meredith. The persuasive effects of direct mail: A regression discontinuity based approach. *Working Paper*, 2010.

R. Gill and J. Robins. Causal Inference for Complex Longitudinal Data: The Continuous Case. *Ann. Statist*, 29(6):1785–1811, 2001.

P. W. Glewwe, N. Ilias, and M. Kremer. Teacher Incentives. *NBER Working Paper*, 2003.

R. Gordon, T. J. Kane, and D. O. Staiger. Identifying effective teachers using performance on the job. *Hamilton Project Discussion Paper 2006-01*, 2006.

J. Hahn, P. Todd, and W. V. der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

E. Hanushek and L. Woessmann. Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *NBER Working Paper*, 2005.

E. A. Hanushek. Conceptual and Empirical Issues in the Estimation of Educational Production Functions. *The Journal of Human Resources*, 14(3):351–388, 1979.

E. A. Hanushek. The failure of input-based schooling policies. *Economic Journal*, 113(485):64–98, 2003.

D. N. Harris and T. R. Sass. Value-Added Models and the Measurement of Teacher Quality. *Unpublished Manuscript: http://myweb.fsu.edu/tsass/*, 2006.

J. Heckman and S. Navarro. Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics*, 136(2):341–396, 2007.

J. Heckman, R. Lalonde, and J. Smith. The economics and econometrics of active labor market programs. *Handbooks in Economics*, (5):1865–2085, 1999.

M. Hernan, B. Brumback, and J. Robins. Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments. *Journal of the American Statistical Association*, 96(454), 2001.

K. Hirano and J. Porter. Asymptotics for statistical treatment rules. *Econometrica*, 2009.

K. Hirano, G. Imbens, D. Rubin, and X. Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, 2000.

N. Hjort and M. Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, pages 1619–1647, 1996.

P. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

T. Holmes. The effect of state policies on the location of manufacturing: Evidence from state borders. *Journal of Political Economy*, 106(4):667–705, 1998.

G. Hong and S. Raudenbush. Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, 33(3):333, 2008.

P. Huber. *Robust statistics*. Wiley New York, 1981.

G. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008.

G. Imbens and D. Rubin. Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. *The Annals of Statistics*, 25(1):305–327, 1997.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62:467–475, 1994.

G. W. Imbens and K. Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *Harvard University: Working Paper*, 2008.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, and in the Social and Biomedical Sciences*. Cambridge University Press, forthcoming.

B. A. Jacob and L. Lefgren. Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, 86(1):226–244, 2004.

B. A. Jacob, L. J. Lefgren, and D. P. Sims. The persistence of teacher-induced learning gains. *Journal of Human Resources*, 45(4):915–943, 2010.

M. Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146, 1993.

K. Kalyanaraman. Bandwidth choice for regression functionals with application to average treatment effects. *Harvard University: Working Paper*, 2009.

T. Kane, S. Riegg, and D. Staiger. School quality, neighborhoods, and housing prices. *American Law and Economics Review*, 8(2):183, 2006.

T. J. Kane and D. O. Staiger. The Promise and Pitfalls of Using Imprecise School Accountability Measures. *The Journal of Economic Perspectives*, 16(4):91–114, 2002.

T. J. Kane and D. O. Staiger. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *NBER Working Paper 14607*, 2008.

M. Kitahata, S. Gange, A. Abraham, B. Merriman, M. Saag, A. Justice, R. Hogg, S. Deeks, J. Eron, J. Brooks, et al. Effect of early versus deferred antiretroviral therapy for HIV on survival. *The New England Journal of Medicine*, 360(18):1815, 2009.

M. Kremer, E. Miguel, and R. L. Thornton. Incentives to Learn. *NBER Working Paper*, 2003.

A. B. Krueger. Economic Considerations and Class Size. *Economic Journal*, 2003.

A. B. Krueger and D. M. Whitmore. The Effect of Attending a Small Class in the Early Grades on College-test Taking and Middle School Test Results: Evidence from Project Star. *The Economic Journal*, 111(468):1–28, 2001.

H. F. Ladd and R. P. Walsh. Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education Review*, 21(1):1–17, 2002.

R. Lalive. How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of Econometrics*, 142(2):785–806, 2008.

A. Lang. When and how should treatment be started in Parkinson disease? *Neurology*, 72(Issue 7, Supplement 2):S39, 2009.

M. Lechner. Sequential causal models for the evaluation of labor market programs. *Journal of Business and Economic Statistics*, 27:71–83, 2009.

D. Lee. Randomized experiments from non-random selection in US House elections. *Journal of Econometrics*, 142(2):675–697, 2008.

D. Lee and D. Card. Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674, 2008.

D. S. L. T. L. D. S. Lee and T. Lemieux. Regression discontinuity designs in economics. *NBER Working Paper 14723*, 2009.

R. Little and D. Rubin. *Statistical analysis with missing data*. Wiley New York, 2 edition, 1987.

C. Loader. Local likelihood density estimation. *The Annals of Statistics*, pages 1602–1618, 1996.

J. Lok, M. Hernan, and J. Robins. Optimal start of treatment based on time–dependent covariates. *Working Paper*, 2008.

F. M. Lord. A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68(5):304–305, 1967.

J. Ludwig and D. Miller. Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design. *The Quarterly Journal of Economics*, 122 (1):159–208, 2007.

C. Manski. Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice. *Journal of Econometrics*, 95(2):415–442, 2000.

C. Manski. Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, 72(4):1221–1246, 2004.

F. Martorell. Do High School Graduation Exams Matter? Evaluating the Effects of Exit Exam Performance on Student Outcomes. *Mimeo, Berkeley Department of Economics*, 2005.

J. Matsudaira. Mandatory summer school and student achievement. *Journal of Econometrics*, 142(2):829–850, 2008.

D. F. McCaffrey. *Evaluating Value-added Models for Teacher Accountability*. Rand Corporation, 2004.

J. McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714, 2008.

W. Mebane and J. Sekhon. Genetic Optimization Using Derivatives: The rgenoud Package for R. *Journal of Statistical Software*, 2009.

R. J. Murnane, J. B. Willett, and F. Levy. The Growing Importance of Cognitive Skills in Wage Determination. *The Review of Economics and Statistics*, 77(2): 251–266, 1995.

S. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B(Statistical Methodology)*, 65(2):331–355, 2003.

S. Murphy, M. van der Laan, and J. Robins. Marginal Mean Models for Dynamic Regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

D. A. Neal and W. R. Johnson. The Role of Premarket Factors in Black-White Wage Differentials. *Journal of Political Economy*, 104(5):869–895, 1996.

W. Newey. Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, pages 233–253, 1994.

J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

J. Papay, R. J. Murnane, and J. B. Willett. The Consequences of High School Exit Examinations for Struggling Low-Income Urban Students: Evidence from Massachusetts. *NBER Working Paper*, 2008.

J. P. Papay, R. J. Murnane, and J. B. Willett. Extending the regression discontinuity approach to multiple assignment variables with an illustration using high school exit examinations. *Working Paper: Harvard University*, 2009.

K. Pence. Foreclosing on opportunity: State laws and mortgage credit. *Review of Economics and Statistics*, 88(1):177–182, 2006.

J. Porter. Estimation in the Regression Discontinuity Model. *Harvard University, unpublished manuscript*, 2003.

S. Raudenbush. Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, 45(1):206, 2008.

S. F. Reardon and J. P. Robinson. Regression discontinuity designs with multiple rating-score variables. *Working Paper*, 2010.

J. Robins. A new approach to causal inference in mortality studies with sustained exposure periods-Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

J. Robins. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health Service Research Methodology: A Focus on AIDS*, 113159, 1989.

J. Robins. Causal inference from complex longitudinal data. *Latent Variable Modeling and Applications to Causality*, 120:69–117, 1997.

J. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, pages 6–10, 1999a.

J. Robins. Association, causation, and marginal structural models. *Synthese*, 121(1): 151–179, 1999b.

J. Robins. Optimal Structural Nested Models for Optimal Sequential Decisions. *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, 2004.

J. Robins and M. Hernan. Estimation of the causal effects of time-varying exposures. *Longitudinal Data Analysis—Fitzmaurice G, Davidian M, Verbeke G, et al, eds*, pages 553–599, 2008.

J. Robins, L. Orellana, and A. Rotnitzky. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27(23), 2008.

J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

J. Rothstein. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1):175–214, February 2010.

D. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34–58, 1978.

D. Rubin. Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test" by D. Basu. *Journal of the American Statistical Association*, 75:591–593, 1980.

D. Rubin. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

D. Rubin. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

D. Rubin. *Statistical Inference for Causal Effects, With Emphasis on Applications in Epidemiology and Medical Statistics*, volume 27, chapter 2, pages 28–63. Elsevier, 2008.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

D. Ruppert and M. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, pages 1346–1370, 1994.

L. J. Schweinhart, J. Montie, Z. Xiang, W. S. Barnett, C. R. Belfield, and M. Nores. *Lifetime effects: The High/Scope Perry Preschool study through age 40*. Ypsilanti, MI: High/Scope Press., 2005.

S. Shavell and L. Weiss. The optimal payment of unemployment insurance benefits over time. *The Journal of Political Economy*, 87(6):1347–1362, 1979.

R. Slavin. Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research*, 57(3):293, 1987.

R. Slavin. Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, 60(3):471, 1990.

Y. Su, A. Gelman, J. Hill, and M. Yajima. Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software*, 20(1): 1–27, 2009.

Y. Sun. Adaptive Estimation of the Regression Discontinuity Model. *Working Paper*, 2005.

D. Thistlethwaite and D. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309, 1960.

P. E. Todd and K. I. Wolpin. On the Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal*, 113(485):3–33, 2003.

J. Tooley and P. Dixon. *Private schools for the poor: A case study from India*. Reading, Royaume-Uni: Centre for British Teachers, 2003.

W. Trochim. *Research design for program evaluation: The regression-discontinuity approach*. Beverly Hills, CA: Sage Publications., 1984.

L. Wasserman. *All of nonparametric statistics*. Springer Science + Business Media, LLC, 2006.

V. C. Wong, P. M. Steiner, and T. D. Cook. Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Working Paper: Northwestern University*, 2010.

L. Yang and R. Tschernig. Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 793–815, 1999.

J. Zhang, D. Rubin, and F. Mealli. Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, 104(485):166–176, 2009.